NOVEL SINGLE AND GENE-BASED TEST PROCEDURES FOR LARGE-SCALE BIVARIATE TIME-TO-EVENT DATA, WITH APPLICATION TO A GENETIC STUDY OF AMD PROGRESSION

by

Yi Liu

BS, Suzhou University, China, 2011

MS, Columbia University, 2013

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

THE GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yi Liu

It was defended on

July 28th 2017

and approved by

Ying Ding, PhD

Assistant Professor

Departmental of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Wei Chen, PhD

Associate Professor

Department of Pediatrics

School of Medicine

University of Pittsburgh

Yu Cheng, PhD

Associate Professor

Departmental of Statistics

Dietrich Dchool of Arts and Sciences

University of Pittsburgh

Daniel E. Weeks, PhD

Professor

Department of Human Genetics

Graduate School of Public Health

University of Pittsburgh

Jong H. Jeong, PhD

Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Dissertation Director: **Ying Ding**, PhD

Assistant Professor

Departmental of Biostatistics

Graduate School of Public Health

University of Pittsburgh

NOVEL SINGLE AND GENE-BASED TEST PROCEDURES FOR LARGE-SCALE BIVARIATE TIME-TO-EVENT DATA, WITH APPLICATION TO A GENETIC STUDY OF AMD PROGRESSION

Yi Liu, PhD

University of Pittsburgh, 2017

Abstract:

Motivated by a genome-wide association study (GWAS) to discover risk variants for the progression of Age-related Macular Degeneration (AMD), I develop a computationally efficient copula-based score test, in which the association between bivariate progression times is explicitly modeled. Specifically, a two-step estimation approach with numerical derivatives to approximate the score function and information matrix is proposed. Both parametric and weakly parametric marginal distributions under the proportional hazards (PH) assumption are considered. Extensive simulation studies are conducted to evaluate the Type I error control and power performance of the proposed method. Further I extend this work to gene-based tests through the functional linear regression approach, which models the variants (within the same gene region) as a function of their physical positions. A robust variance estimator for bivariate time-to-event data under functional linear model is also proposed. Simulation studies are conducted to evaluate the Type I error control and power performance of the proposed method. Finally, we apply our method on a large randomized trial data set, Age-related Eye Disease Study (AREDS), to identify progression risk variants and gene regions for AMD progression. The top variants identified in the ARMS2 gene on chromosome 10 show differential progression profiles for different genetic groups, which are useful in characterizing and predicting the risk of progression for patients with moderate AMD.

Public health significance: The application of the proposed methods jointly models the progression profiles in both eyes, which has not been done in any of the previous studies of AMD progression. The findings provide new insights about the genetic causes on AMD progression from single variants to genes, which will be critical for establishing novel and reliable predictive models of AMD progression to accurately identify high-risk patients at an early stage.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 MOTIVATING EXAMPLE

Our research is motivated by a genome-wide association study (GWAS) on identifying risk variants for progression of a bilateral eye disease – Age-related Macular Degeneration (AMD). AMD is a common, polygenic, and progressive neurodegenerative disease, which is a leading cause of blindness in the developed world (Swaroop et al., 2009; The Eye Diseases Prevalence Research Group, 2004). The overall estimated prevalence of any AMD was 6.5% in the United States, with a 95% confidence interval $[5.5\%, 7.6\%]$, and the prevalence of late AMD was 0.8%, with a 95% confidence interval $[0.5\%, 1.3\%]$ (Klein et al., 2011).

The age-related eye disease study (AREDS) was a multi-center, controlled, randomized clinical trial of AMD sponsored by National Eye Institute. It was designed to assess the clinical course and risk factors for the development and progression of AMD. In this cohort, participants were longitudinally followed up to 12 years to examine the progression of the disease (Age-Related Eye Disease Study Research Group, 1999). The study collected DNA samples of consenting participants and performed genome-wide genotyping. The objective of our study is to discover risk variants and genes for AMD progression using the AREDS dataset.

1.2 GENOME-WIDE ASSOCIATION STUDY

With wide-spread availability of high-throughput genotyping technology, large scale GWAS become a powerful tool to discover risk variants for complex diseases. One ultimate goal of

these association studies is to uncover the biological underpinnings of disease susceptibility (Wang et al., 2005). Results of such studies can subsequently lead to better understanding of the disease process and the development of improved strategies for disease prevention and treatment.

The most common GWAS approach is based on case-control samples, which compares a group of diseased individuals with healthy individuals to discover significant variants associated with the disease (Wu et al., 2010; The Wellcome Trust Case Control Consortium, 2007). Another popular GWAS approach studies the quantitative traits of individuals such as gene expression or biomarker concentration (Cho et al., 2009). More recently, GWAS on time-to-event data have become increasingly popular to study the progression or survival profiles of certain diseases. For example, Azzato et al. (2010) performed a GWAS of survival after diagnosis of breast cancer. Pillas et al. (2010) conducted a GWAS for time to first tooth eruption. Ioannidis et al. (2010) did a thorough review of cancer-related GWAS.

1.3 GENETIC STUDIES OF AMD PROGRESSION

Both common and rare variants associated with AMD risk (i.e., whether or not to develop the disease) have been identified in multiple large-scale case-control association studies (Chen et al., 2010a; Fritsche et al., 2013, 2016). However, the genetic causes of AMD progression have not been well studied.

Recently, several studies evaluated the effects of a few known AMD risk variants on its disease progression (Seddon et al., 2009, 2014). These studies analyzed only one eye per subject (e.g., the faster progressed eye). More recently, Sardell et al. (2016) and Ding et al. (2017) evaluated a set of known AMD risk variants on progression using both eyes with a robust marginal Cox model, where the between-eye correlation was taken into account. All the existing studies on AMD progression only analyzed a small set of known AMD risk variants. To the best of our knowledge, no large-scale studies have been performed to discover the variants associated with AMD progression.

To understand the genetic underpinning of progression of this bilateral disease, our first objective is to develop a computationally efficient test procedure for bivariate time-to-event data to perform GWAS for single markers. The second objective is to extend this single marker test procedure to gene-based association analysis.

2.0 BACKGROUND

2.1 TIME-TO-EVENT DATA ANALYSIS

Time-to-event data analysis, also known as survival analysis, is a branch of statistics for analyzing the expected duration of time for an event to happen. In life science applications, such events can be death, disease/tumor progression, and etc.

Define a random variable T as the time from the beginning of an observation period to the occurrence of an event. The survival function $S(t)$ describes the probability of event happens prior to time t. Specific to time-to-event analysis, when the follow-up time is not long enough to capture the event, or patients drop off during the follow-up, call it censored. Effects of different censoring schemes have been widely studied. In this work, we consider the right censored scenario when censoring time is independent of event time. Define a random variable for censoring time as C and censoring indicator as $\Delta = I(T \leq C)$. The observed time is denoted by $Y = \min(T, C)$. The observed data are

$$D = \{(Y, \Delta) : Y = \min(T, C), \Delta = I(T \leq C)\}.$$

The cumulative hazard $H(t)$ and the instantaneous hazard $\lambda(t)$ are two useful functions in survival analysis, where $S(t) = \exp(-\Lambda(t)) = \exp(-\int \lambda(t)dt)$. In time-to-event data analysis, the most commonly used approach is the Cox proportional hazards model (Cox, 1972) which models the instantaneous hazards as a regression function of covariates. Another commonly used approach is the accelerated failure time model (Wei, 1992), which directly models the transformed survival time as a function of covariates.

2.2 METHODS FOR BIVARIATE TIME-TO-EVENT DATA

The standard techniques to analyze time-to-event data are based on the assumption that event times are independent of each other. However, this assumption can be violated when the study units are paired such as twins, married couples or bilateral objects such as eyes in ophthalmology studies. In the presence of the dependence between the event times, multivariate survival analysis needs to be considered. Hougaard (2000) and Joe (1997) provide thorough reviews and examples for multivariate survival analysis.

First, we introduce notation for bivariate time-to-event data. Assume there are n subjects. Let (T_{1i}, T_{2i}) and $(C_{1i}, C_{2i}), i = 1, ..., n$, denote the bivariate failure times and censoring times, respectively. Denote by $X = (X_{1i}, X_{2i})$ the risk factors for the ith subject. Assume given the covariates X, (T_{1i}, T_{2i}) and (C_{1i}, C_{2i}) are independent. Then for each subject, we observe

$$D_i = \{(Y_{1i}, Y_{2i}, \Delta_{1i}, \Delta_{2i}, X_{1i}, X_{2i}) : Y_{ki} = \min(T_{ki}, C_{ki}), \Delta_{ki} = I(T_{ki} \leq C_{ki}), k = 1, 2\}.$$

Let $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ denote the joint survival function for the bivariate failure time (T_1, T_2) and let $f(t_1, t_2)$ denote its corresponding density function. Denote by θ all the parameters in $S(t_1, t_2)$, then the joint likelihood for the observed data $\{D_i\}_{i=1}^n$ can be written as

$$\begin{aligned} L(\theta; D = (Y_1, Y_2, \Delta_1, \Delta_2, X_1, X_2)) \\ = \prod_{i=1}^{n} f(y_{1i}, y_{2i}|x_{1i}, x_{2i})^{\delta_{1i}\delta_{2i}} \times \left[-\frac{\partial S(y_{1i}, y_{2i}|x_{1i}, x_{2i})}{\partial y_{1i}}\right]^{\delta_{1i}(1-\delta_{2i})} \\ \times \left[-\frac{\partial S(y_{1i}, y_{2i}|x_{1i}, x_{2i})}{\partial y_{2i}}\right]^{(1-\delta_{1i})\delta_{2i}} \times S(y_{1i}, y_{2i}|x_{1i}, x_{2i})^{(1-\delta_{1i})(1-\delta_{2i})}, \end{aligned} \tag{2.1}$$

where $(\delta_{1i}, \delta_{2i}) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$.

2.2.1 Marginal model

The first approach for analyzing bivariate time-to-event data is a marginal method, which was developed under the General Estimation Equation (GEE) framework. A robust sandwich estimator from the estimating equation is used to estimate the variance-covariance matrix of the regression parameter. For example, Wei et al. (1989) considered the semi-parametric Cox model and proposed to estimate the regression parameter β under a working independence assumption by which observations in each cluster are treated as independent of one another.

Assume the regression coefficients are the same across each margin. Define for each margin k,

$$Y_{ki}(t) = I\{T_{ki} \geq t\}.$$

Then, under the Cox proportional hazards assumption, the bivariate partial score function for β can be written as

$$\begin{aligned} U(\beta) &= \sum_{k=1}^{2}\sum_{i=1}^{n} U_{ki}(\beta) \\ &= \sum_{k=1}^{2}\sum_{i=1}^{n} \Delta_{ki}\left\{X_{ki} - \frac{S_k^{(1)}(t_{ki};\beta)}{S_k^{(0)}(t_{ki};\beta)}\right\}, \end{aligned} \tag{2.2}$$

where $S_k^{(0)}(t;\beta) = \sum_{i=1}^{n} Y_{ki}(t_{ki})\exp(\beta X_{ki})$ and $S_k^{(1)}(t;\beta) = \sum_{i=1}^{n} Y_{ki}(t)\exp(\beta X_{ki})X_{ki}$, $k = 1, 2$. Define $\hat{\beta}$ to be the root of $U(\beta) = 0$. If the marginal Cox regression model is correctly specified, $n^{-\frac{1}{2}}U(\beta)$ is asymptotically normally distributed with mean zero and variance

$$B = E[U_{ki}^2(\beta)],$$

which, in practice, can be estimated by

$$\hat{B} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n}[U_{ki}^2(\beta)]|_{\beta=\hat{\beta}}. \tag{2.3}$$

In another form, variance of $n^{-\frac{1}{2}}U(\beta)$ can expressed by

$$A = -E[\partial U_{ki}(\beta)/\partial\beta]. \tag{2.4}$$

Note that (2.4) can be consistently estimated by

$$\hat{A} = -\frac{1}{2n}\sum_{k=1}^{2}\sum_{i=1}^{n} \partial U_{ki}(\beta)/\partial\beta|_{\beta=\hat{\beta}},$$

and robust inference is based on the variance $\hat{\Gamma} = \hat{A}^{-1}\hat{B}(\hat{A}^{-1})'$. It is closely related to the GEE methodology (Liang and Zeger, 1986). Lee et al. (1992) showed that $\hat{\beta}$ was consistent with $n^{-\frac{1}{2}}(\hat{\beta} - \beta) \to N(0, \Gamma)$, where $\Gamma = A^{-1}B(A^{-1})'$.

A marginal model is useful when the main interest is to estimate the effect of covariates on survival. By applying a "sandwich" estimator, it takes into account the fact that observed event times are correlated. However, the strength of such correlations cannot be explicitly modeled under this marginal approach.

2.2.2 Copula model

One of the earliest distribution families for modelling correlated bivariate measurements is the copula family (Clayton, 1978), originated from Sklar's Theorem (Sklar, 1959), in which the joint distribution is modeled as a function of each marginal distribution together with an association parameter. Copula function provides a parametric assumption about the association between two correlated margins. A bivariate copula is a function defined as $\{C_\eta : [0,1]^2 \to [0,1] : (u,v) \to C_\eta(u,v), \eta \in R\}$. Assume U and V are both uniformly distributed random variables. The parameter η in the copula function describes the dependence between U and V. By Sklar's theorem (Sklar, 1959), one can model the joint distribution by modeling the copula function and the marginal distributions separately. The theorem is stated as: if marginal survival functions $S_1(t_1) = P(T_1 > t_1)$ and $S_2(t_2) = P(T_2 > t_2)$ for T_1 and T_2 are continuous, then there exists a unique copula function C_η such that for all $t_1 \geq 0, t_2 \geq 0$,

$$S(t_1, t_2) = C_\eta((S_1(t_1), S_2(t_2)),\ t_1, t_2 \geq 0.$$

Define the density function for C_η to be $c_\eta = \partial^2 C_\eta(u,v)/\partial\, u\partial\, v$, then the joint density function of T_1 and T_2 can be expressed as

$$f(t_1, t_2) = c_\eta(S_1(t_1), S_2(t_2))f_1(t_1)f_2(t_2),\ t_1, t_2 \geq 0.$$

The Copula function is robust in modeling various dependence structures and has nice properties. For example, the rank-based dependence measurement Kendall's τ can be directly obtained as a function of η in some copula models.

In this work, we focus on the Archimedean copula family, which is one of the most popular copula families because of its flexibility and simplicity. A copula C_η belongs to an Archimedean family if it can be expressed as:

$$C_\eta(u, v) = H(H^{-1}(u; \eta) + H^{-1}(v; \eta)),\ 0 \leq u, v \leq 1.$$

$H : [0, \infty) \to [0, 1]$ is the so-called generating function for Archimedean copulas. It is continuous, strictly decreasing and convex satisfying $H(0; \eta) = 0$.

Two most frequently used Archimedean copulas in survival analysis are:

Clayton copula (Clayton, 1978)

$$C_\eta(u, v) = (u^{-\eta} + v^{-\eta} - 1)^{-1/\eta},\ \eta \in (0, \infty),$$

and

Gumbel-Hougaard copula (Gumbel, 1960)

$$C_\eta(u, v) = \exp\{-[(-\log u)^\eta + (-\log v)^\eta]^{1/\eta}\},\ \eta \in [1, \infty).$$

The Clayton copula models lower tail dependence in survival functions, while a Gumbel copula models upper tail dependence in survival functions. For a Clayton copula, the association parameter η corresponds to Kendall's τ as $\tau = \frac{\eta}{\eta+2}$. Thus, T_1 and T_2 are positively associated when $\eta > 0$ and are independent when $\eta \to 0$. While for a Gumbel copula, $\tau = \frac{\eta-1}{\eta}$, meaning T_1 and T_2 are positively associated when $\eta > 1$ and are independent when $\eta = 1$. Figure 1 shows a visual example of how margins are correlated under the Clayton copula with Weibull marginal distributions. The figure indicates that as Kendall's τ increases, both (T_1, T_2) and (S_1, S_2) become more linearly related.

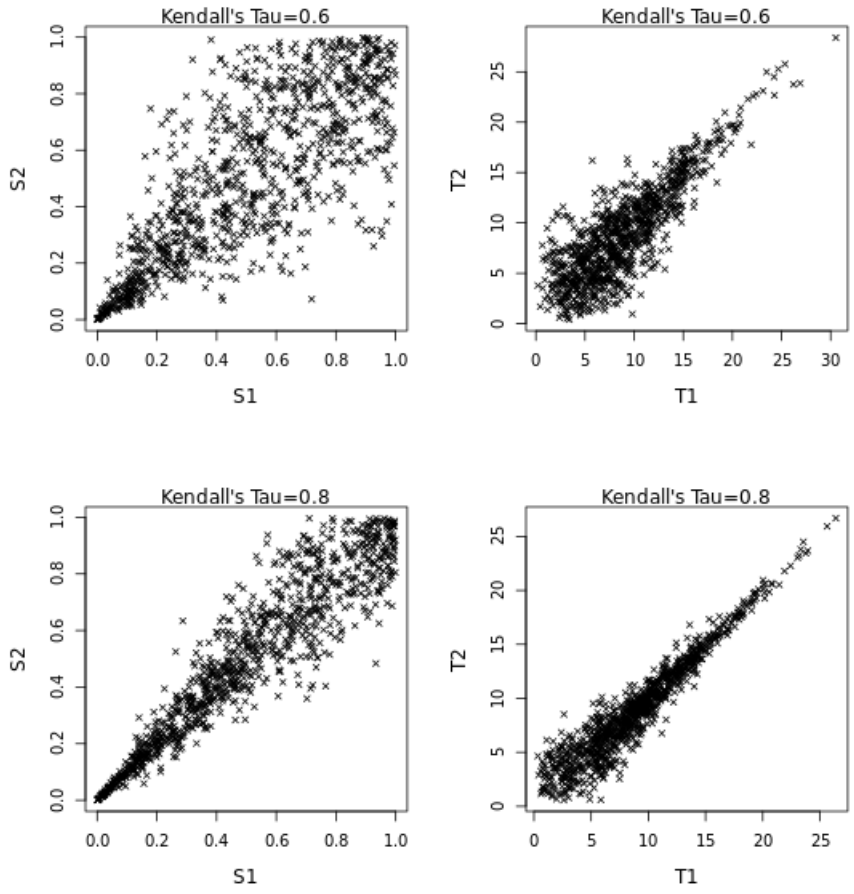

Figure 1: Marginal association on survival probability and event times under Clayton copula with Weibull margins (no censoring).

S_k, T_k, $k = 1, 2$ are the survival functions and observed event times respectively.

Under the copula model, the joint likelihood function (2.1) can be rewritten as

$$\begin{aligned} L((\eta, S_1, S_2); D) = \prod_{i=1}^{n} & [c_\eta(S_1(y_{1i}|x_{1i}), S_2(y_{2i}|x_{2i}) f_1(y_{1i}|x_{1i}) f_2(y_{2i}|x_{2i})]^{\delta_{1i}\delta_{2i}} \\ & \times \left[-\frac{\partial\, C_\eta(S_1(y_{1i}|x_{1i}), S_2(y_{2i}|x_{2i}))}{\partial y_{1i}} \right]^{\delta_{1i}(1-\delta_{2i})} \\ & \times \left[-\frac{\partial\, C_\eta(S_1(y_{1i}|x_{1i}), S_2(y_{2i}|x_{2i}))}{\partial y_{2i}} \right]^{(1-\delta_{1i})\delta_{2i}} \\ & \times C_\eta(S_1(y_{1i}|x_{1i}), S_2(y_{2i}|x_{2i}))^{(1-\delta_{1i})(1-\delta_{2i})}. \end{aligned} \tag{2.5}$$

One important characteristics for Clayton copula model is that the ratio of two conditional hazards between T_1 given $T_2 = t_2$ and T_1 given $T_2 \geq t_2$ is a constant, namely,

$$\frac{\lambda_1(t_1|T_2 = t_2)}{\lambda_1(t_1|T_2 \geq t_2)} = 1 + \eta.$$

This ratio is also known as cross ratio (Clayton, 1978; Oakes, 1982), a popular dependence measurement for multivariate survival data in addition to Spearman's correlation and Kendall's τ.

2.2.3 Frailty model

Another popular approach for analyzing multivariate survival data is the frailty model, which was originally proposed by Oakes (1982) and Vaupel et al. (1979). In this approach, a common latent frailty variable, as a random effect, introduces the correlation between survival times under Cox Proportional hazards model. The hazard function for each margin for a given subject can be specified as,

$$\lambda_{i,1}(t_{i,1}; \beta_1|w_i) = w_i\lambda_1(t_{i,1}; \beta_1)$$
$$\lambda_{i,2}(t_{i,2}; \beta_2|w_i) = w_i\lambda_2(t_{i,2}; \beta_2),$$

where $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ are hazard functions for each margin. Here the frailty random variable $w_i \sim f_\rho(w)$ follows a prespecified distribution with parameter ρ. For example, Gamma and positive stable distributions are commonly used distributions in practice. To obtain the joint survival function, one can derive the two marginal survival functions by integrating out the random effect,

$$S(t_1; \beta_1) = \int_0^\infty e^{-w\Lambda_1(t_1;\beta_1)} f_\rho(w)dw = \mathcal{L}_p(\Lambda_1(t_1; \beta_1))$$
$$S(t_2; \beta_2) = \int_0^\infty e^{-w\Lambda_2(t_2;\beta_2)} f_\rho(w)dw = \mathcal{L}_p(\Lambda_2(t_2; \beta_2)). \tag{2.6}$$

Here $\mathcal{L}_p(\cdot)$ is the Laplace transformation on the density function of the random frailty term w.

Assume the conditional independence of T_1 and T_2 given w, one can express the joint distribution of T_1 and T_2 as

$$S(t_1, t_2; \beta_1, \beta_2|w) = S_1(t_1; \beta_1|w)S_2(t_2; \beta_2|w). \tag{2.7}$$

Combining (2.6) and (2.7) the joint survival function can be explicitly written out as:

$$\begin{aligned} S(t_1, t_2; \beta_1, \beta_2) &= \int_0^\infty S_1(t_1; \beta_1|w)S_2(t_2; \beta_2|w)f_\rho(w)dw \\ &= \int_0^\infty e^{-w(\Lambda_1(t_1;\beta_1)+\Lambda_2(t_2;\beta_2))} f_\rho(w)dw \\ &= \mathcal{L}_\rho[\mathcal{L}_\rho^{-1}(S_1(t_1; \beta_1)) + \mathcal{L}_\rho^{-1}(S_2(t_2; \beta_2))]. \end{aligned} \tag{2.8}$$

2.2.4 Relationship between Archimedean copula model and frailty model

The marginal Archimedean copula model, $C_\alpha(u, v) = H(H^{-1}(u; \eta) + H^{-1}(v; \eta))$ and the shared frailty model (2.8) are two approaches to model bivariate survival function $P(T_1 > t_1, T_2 > t_2)$. By integrating out the frailty variable, the joint survival function ends up being the same as an Archimedean copula. There are some equivalencies between the two general approaches.

From section 2.2.3, the joint survival function is derived as $S(t_1, t_2) = \mathcal{L}(\Lambda_1(t_1) + \Lambda_2(t_2))$ with $\Lambda_k(t) = \mathcal{L}^{-1}(S_k(t))$. Thus one can write out the joint survival function as $S(t_1, t_2) = \mathcal{L}[\mathcal{L}^{-1}(S_1(t_1)) + \mathcal{L}^{-1}(S_2(t_2))]$. This form is the same as the form of an Archimedean copula with a generating function equal to $\mathcal{L}^{-1}$.

For example, we compare the Gamma frailty model and Clayton copula. The density function for a one-parameter Gamma distribution with mean 1 and variance η is

$$f_U(u) = \frac{u^{((1/\eta)-1)} \exp(-u/\eta)}{\eta^{1/\eta}\Gamma(1/\eta)}.$$

The Laplace transformation of the Gamma frailty density is

$$\mathcal{L}(s) = (1 + \eta s)^{-1/\eta}$$

and

$$\mathcal{L}^{-1}(s) = \frac{s^{-\eta} - 1}{\eta}, \eta \geq 0.$$

Pluging into (2.8) the joint survival function becomes:

$$S(t_1, t_2) = [S_1(t_1)^{-\eta} + S_2(t_2)^{-\eta} - 1]^{-1/\eta}.$$

This is the same as the Clayton copula. A Similar relationship can be derived for Gumbel copula and positive stable frailty model with generating function $H^{-1}(s) = L(s) = \exp(-s^{1/\eta})$.

For the estimation procedure, a two-stage approach proposed by Shih and Louis (1995) are often used for solving a copula model. The frailty model is solved based on conditional likelihoods. This may lead to different parameter estimates in real applications.

2.3 SINGLE MARKER TESTS IN GENETIC STUDIES

In typical genome-wide association studies, single marker tests are performed in most of scenes. Such tests examine the association between a single nucleotide polymorphism (SNP) and a trait once at a time. Among choices of single marker tests, the score test is usually preferred than other types of likelihood-based tests, like the Wald test or the likelihood ratio test (LRT) (Cantor et al., 2010; Sha et al., 2011). This is because in large scale screening, the score test only needs to fit the model under the null (i.e., no SNP effect) rather than fitting millions of (alternative) models, one for each SNP. Because of this feature, the score test is computationally more efficient and stable compared to the Wald test and the LRT. There are many readily available computer programs for conducting single marker GWAS on quantitative, binary, count and censored traits.

2.4 GENE-BASED TESTS FOR TIME-TO-EVENT DATA

In genome-wide association studies, single variant test is powerful in detecting possible signals across whole genome for common variants and variant with relatively large effect sizes. However, there are some limitations when applying single variant test. For example, the single variant test lacks of power to detect association with rare variants or variants with

small effect sizes. In addition, most of the statistical approaches used for testing single variants are not stable when the minor allele frequency is small. Therefore, there is an increasing interest to develop new statistical methods to investigate the relationship between rare variants and disease susceptibility (Wu et al., 2010; Lin et al., 2011; Gorlov et al., 2008).

2.4.1 Burden test

Burden test was first proposed by Li and Leal (2008) for binary traits for the purpose of detecting association with rare variants for the common diseases. The general idea of the burden test for genetic studies is based on collapsing rare variants in a genetic region to be a summary variable, which is then used for testing the association with the phenotypes. Later Han and Pan (2010) extended the work for censored traits under Cox proportional hazards framework. The burden test has been demonstrated to be most powerful when all causal variants have the same effect direction.

Define X_i to be non-genetic covariates and G_i to be the genotype for individual i. Assume one is interested in testing the overall effect of m variants in a region. For censored traits, under Cox proportional hazards assumption, the burden test can be formulated as:

$$\lambda_i(t) = \lambda_0(t)e^{X_i\beta + G_iW_i\gamma},$$

where W_i is a pre-defined $m \times m$ diagonal weight matrix (e.g. a function of minor allele frequency). β and γ are regression coefficients for non-genetic and genetic effects respectively.

Assume the effect sizes for all variants are homogeneous with $\gamma = \gamma_0$, then the hypothesis for testing the genetic effects can be written as:

$$H_0 : \gamma_0 = 0 \qquad vs. \qquad H_1 : \gamma_0 \neq 0,$$

which is a burden test with the collapsed genetic burden score $\sum_{j=1}^{m} G_{ij}W_{jj}$, where G_{ij} is the jth element of vector G_i for individual i, and W_{jj} is the j the diagonal element of the weight matrix W. All types of likelihood-based test like Wald, LRT and score test can be performed.

2.4.2 Sequence kernel association test (SKAT)

The sequence kernel association test (SKAT) was proposed by Wu et al. (2011), is usually considered as a computationally efficient score test on variance component parameter to test for association between genetic variants in a region and different types of traits. One of the advantages of SKAT is it can quickly calculate p-values by fitting the null model containing only the non-genetic covariates. When the linear kernel is used, the test statistic can be expressed as a weighted sum of the statistics from single marker score test.

In 2014, Chen et al. (2014) extended SKAT to survival traits. With survival phenotypes modeled by the Cox PH model, the hazard function can be written out as:

$$\lambda_i(t) = \lambda_0(t)e^{X_i\beta+G_iW\gamma},$$

where β is p fixed effects for the non-genetic covariates, γ is m random effects with mean 0 and variance $\sigma^2 I_m$ for the genotypes. W is a pre-define weight matrix. The SKAT formulates the test hypothesis as:

$$H_0 : \sigma^2 = 0 \qquad vs. \qquad H_1 : \sigma^2 > 0.$$

The SKAT statistic is

$$Q = r'GWWG'r \sim \sum_{j=1}^{m} \lambda_j \chi^2_{1,j},$$

where Σ is the covariance matrix of the vector $WG'r$, r is the vector of martingale residuals calculated from the null model, λ_j are the eigenvalues of Σ, and $\chi^2_{1,j}$ are independent χ^2 distributions with degree of freedom 1. Wu et al. (2011); Lin and Tang (2011) demonstrated that SKAT is superior to other existing rare variant tests, especially in the presence of both protective and detrimental rare variants.

2.4.3 Functional regression for time-to-event data

A limitation of the burden test and SKAT is the lack of effectively utilizing linkage disequilibrium (LD) or in other words, correlation among genetic variants, effectively. Fan et al. (2013) first introduced the idea of functional regression (FR) for testing associations between

genetic variants and quantitative traits. The FR-based model treats the contribution of genetic markers as a function of physical positions and a realization of a stochastic process. Recently, Fan et al. (2016) extended this work to censored traits and applied the method on AREDS dataset using only left eye information.

Assume in a genomic region, m variants are sequenced with ordered physical positions, e.g., base pair positions, $0 \leq u_1 < ... < u_m$. Let $G_i = (g_i(u_1), ..., g_i(u_m))'$ denote the genotype information and $X_i = (x_{i1}, ..., x_{ip})$ denote a $p \times 1$ vector of fixed effect of non-genetic covariates for subject i. A FR-based Cox proportional hazards model can be written as:

$$\lambda_i(t) = \lambda_0(t) \exp\left(X_i'\beta + \int_0^1 G_i(u)\gamma(u)du\right), \tag{2.9}$$

where $\lambda_0(s)$ is the baseline hazard function, and β is a $p \times 1$ vector of coefficients for fixed non-genetic effect covariates. $G_i(u)$ is the genetic variant function (GVF) and $\gamma(u)$ is the genetic effect function (GEF) at the position u. In practice, one can approximate the integration part in (2.9) by a summation term

$$\lambda_i(t) = \lambda_0(t) \exp\left(X_i'\beta + \sum_{j=1}^{m} g_i(u_j)\gamma(u_j)\right). \tag{2.10}$$

In both representation of (2.9) and (2.10), $\gamma(u)$ is assumed to be smooth and can be approximated using smoothing techniques like B-spline or Fourier spline (Ramsay et al., 2009). Define a series of B_γ basis function by $\psi(u) = (\psi_1(u), ...\psi_{B_\gamma}(u))'$ and a $B_\gamma \times 1$ coefficient vector $(\gamma_1, ..., \gamma_{B_\gamma})'$, then $\gamma(u)$ can be approximated by

$$\hat{\gamma}(u) = (\psi_1(u), ...\psi_{B_\gamma}(u))(\gamma_1, ..., \gamma_{B_\gamma})'. \tag{2.11}$$

A standard test procedure to test whether the variants in a region are associated with the outcome can be translated to test the null hypothesis:

$$H_0 : \gamma_1 = ... = \gamma_{B_\gamma} = 0.$$

With a model specified using such a smoothing technique, instead of estimating m parameters, we have only B_γ ($\ll$ m in most cases) parameters to estimate.

3.0 A COPULA-BASED SCORE TEST ON SINGLE MARKERS FOR BIVARIATE TIME-TO-EVENT DATA

3.1 BACKGROUND

As discussed in section 1.2, fast development of high-throughput genotyping technology makes single variant tests for each SNP widely applied in detecting genome wide associations for disease phenotypes during the past decade. From the most commonly used case-control setup to quantitative phenotypes, many statistical methods have been established. Motivated by the study of identifying risk variants associated with AMD progression, with progression times available for both eyes, a test procedure suitable for large-scale bivariate time-to-event data is in need.

Three major approaches for dealing with bivariate censored data are reviewed in section 2.2 for their pros and cons. Given the objective of our study is to discover risk variants for the progression of this bilateral disease, we propose to develop a test procedure based on copula models, so that we can (1) assess the genetic effect on a marginal (population) level, and (2) explicitly model the correlation strength while accommodating a very large number of "clusters" (i.e., subjects).

In the GWAS setting, the score test is usually preferred, given it usually requires less computing time, especially when the model fitting is time consuming. In this work, we develop a computationally efficient copula-based score test procedure for analyzing bivariate time-to-event data, and then apply it on AREDS data to identify significant variants associated with AMD progression.

3.2 SCORE TEST UNDER COPULA FRAMEWORK

We consider testing each single SNP in a GWAS setting. Specifically, we are interested in testing the null hypothesis that, whether a given SNP is associated with disease progression, after adjusting for other risk factors. In this work, we consider the marginal distributions under the Cox PH assumption. We then further denote by $S_0 = (S_{01}, S_{02})$ the baseline survival functions for T_1 and T_2, and $\beta = (\beta_g, \beta_{ng})$ the regression coefficients, where β_{ng} are the coefficients of non-genetic risk factors and β_g is the coefficient of the SNP. In this work, we assume the regression coefficients β are the same for T_1 and T_2, which is scientifically plausible for the bilateral eye disease we consider here. However, the method can be easily generalized to the situation where each T_k has its own regression coefficients.

Denote by a $p \times 1$ vector $\theta_{p\times 1} = (\beta', \alpha', \eta)$ the full parameter set for the copula model, where $\beta' = (\beta_g, \beta_{ng})'$ denotes the regression coefficients, $\alpha' = (\alpha_1, \alpha_2)'$ denotes the parameters in $S_{0k}(\cdot)$, η is the association parameter. We are interested in testing whether or not $\beta_g = 0$. Thus we further separate θ into two parts: $\theta_1 = \beta_g$, which is the scalar parameter of interest (to be tested), and $\theta_2 = (\beta'_{ng}, \alpha', \eta)_{(p-1)\times 1}$, which is the $(p-1)\times 1$ nuisance parameter. Then the null hypothesis can be expressed as $H_0 : \theta_1 = \beta_g = 0$ and θ_2 is arbitrary.

The biggest advantage of the score test in a GWAS setting is that one only need to estimate the parameters once (under the null hypothesis), which is much less computationally intensive as compared to likelihood ratio test or Wald test. This is because all the non-genetic risk factors are the same for testing any of the SNPs. In addition, when minor allele frequency (MAF) of the testing SNP is low, maximizing the complex log-likelihood under a copula model (to obtain the parameter estimates) may produce unstable results. Therefore, we propose to use score test for our problem.

Assume $\hat{\theta}_0 = (\theta_1 = 0, \theta_2 = \hat{\theta}_{20})$ is the restricted maximum likelihood estimate (MLE) of θ from (2.5) under the restriction $\theta_1 = 0$, then the corresponding score function and Fisher's information can be written as

$$U(\hat{\theta}_0) = \frac{\partial}{\partial \theta} \log L(D;\theta)\Big|_{\theta=\hat{\theta}_0} = (U_1'(\hat{\theta}_0), U_2'(\hat{\theta}_0))' = (U_1'(\hat{\theta}_0), 0')',$$

where $U_k(\cdot) = \partial \log L / \partial \theta_k,\ k = 1, 2$, and

$$\mathcal{I}(\hat{\theta}_0) = -E\left[\frac{\partial^2}{\partial \theta^T \partial \theta} \log L(D;\theta)\right]\Big|_{\theta=\hat{\theta}_0} = \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{bmatrix},$$

where $\mathcal{I}_{11}, \mathcal{I}_{12}, \mathcal{I}_{21}$ and $\mathcal{I}_{22}$ are partitions of the information matrix $\mathcal{I}$ by θ_1 and θ_2. Specifically, for single marker test, $\mathcal{I}_{11}$ is a scalar. By Cox and Hinkley (1979), we can obtain the score test statistic as

$$\begin{aligned} Q_s &= U'(\hat{\theta}_0)\mathcal{I}^{-1}(\hat{\theta}_0)U(\hat{\theta}_0) \\ &= (U_1'(\hat{\theta}_0), 0')\mathcal{I}^{-1}(\hat{\theta}_0)(U_1'(\hat{\theta}_0), 0')' \\ &= U_1'(\hat{\theta}_0)\mathcal{I}^{11}(\hat{\theta}_0)U_1(\hat{\theta}_0), \end{aligned}$$

where $\mathcal{I}^{11} = (\mathcal{I}^{-1})_{11} = 1/(\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21})$.

In practice, the observed information matrix $\mathcal{J}(\hat{\theta}_0)$, where $\mathcal{J}(\theta) = -\frac{\partial^2 \log L(\theta;D)}{\partial \theta' \partial \theta}$, is often used in the score test. With bivariate copula models, the first and second order derivatives of $\log L(D;\theta)$ usually have very complex forms and the forms depend on the specific copula model as well as the marginal distributions. Thus, we propose to use numerical differentiation through Richardson's extrapolation (Richardson, 1911) to approximate the score function and observed information matrix, denoted by $\tilde{U}$ and $\tilde{\mathcal{J}}$. This numerical approximation only requires a close-formed log-likelihood function. Therefore, the score test statistic we propose is

$$\tilde{Q}_s = \tilde{U}^T(\hat{\theta}_0)\tilde{\mathcal{J}}^{-1}(\hat{\theta}_0)\tilde{U}(\hat{\theta}_0) = \tilde{U}_1^T(\hat{\theta}_0)\tilde{\mathcal{J}}^{11}(\hat{\theta}_0)\tilde{U}_1(\hat{\theta}_0), \tag{3.1}$$

which asymptotically follows a χ_1^2 distribution under the null. If $P(\chi_1^2 > \tilde{Q}_s) < 0.05$, we reject H_0 at the 5% level.

3.2.1 Choices of marginal distributions

In this work, we assume the marginal distributions are from the PH family, which can be written as

$$S_k(t_{ki}|x_{ki}) = P(T_{ki} \geq t_{ki}|x_{ki}) = S_{0k}(t_{ki})^{\exp(x_{ki}\beta_k)},\ k = 1, 2,\ i = 1, ..., n.$$

The corresponding hazard function for T_{ki} given covariate X_{ki} can be expressed as

$$\lambda_k(t_{ki}|x_{ki}) = \lambda_{0k}(t_{ki})\exp(x_{ki}\beta_k),\ k = 1, 2,\ i = 1, ..., n, \tag{3.2}$$

where $\lambda_{0k}(t_{ki})$ is the baseline hazard function for the kth survival time. First, we consider both parametric and weakly parametric assumptions for $\lambda_{0k}(\cdot)$. For example, with Weibull distribution,

$$\lambda_{0k}(t) = \gamma_k\lambda_k(\lambda_k t)^{\gamma_k - 1},\ \gamma_k > 0, \lambda_k > 0, k = 1, 2,$$

or with Gompertz distribution,

$$\lambda_{0k}(t) = \gamma_k\lambda_k e^{\lambda_k t},\ \gamma_k > 0, \lambda_k > 0, k = 1, 2,$$

where λ_k is the scale parameter and γ_k is the shape parameter for each baseline margin. In this case, the full parameter set θ is $(\beta_g, \beta_{ng}, \gamma_k, \lambda_k, \eta)$.

In some circumstances, a specific parametric marginal distribution may not fit the data properly. Kim et al. (2007) has shown that the association parameter estimation in copula models is not robust to misspecification of the marginal distributions. Thus, a relaxed weakly parametric assumption such as the piecewise constant hazards may be more desired for marginal distributions. For example,

$$\lambda_{0k}(t) = \rho_{jk} \text{ for } t \in A_{jk} = (a_{(j-1)k}, a_{jk}],\ j = 1, ..., r,\ k = 1, 2,$$

where $0 = a_{0k} < a_{1k} < ... < a_{rk} = \max y_{ik}$ are pre-specified cutoff points. The full parameter set θ in this case will be $= (\beta_g, \beta_{ng}, \rho_{jk}, \eta)$.

In addition to parametric and weakly parametric model for marginal distribution, non-parametric or semiparametric estimates of baseline hazard function are sometimes desired.

For example, one can consider the Breslow estimator (Breslow, 1972) to treat $\lambda_0(\cdot)$ as piecewise constants between all uncensored failure times. A pseudo-maximum likelihood (PML) estimation can be implemented by fixing the cumulative baseline hazard $\Lambda_{0k}(t)$ with its estimate from the marginal model.

For example, the Breslow estimator proposed by Breslow (1972) can be used by treating $\lambda_0(\cdot)$ as piecewise constant between uncensored failure times in the proportional hazards model. In such a case, the Breslow estimator is

$$\hat{\Lambda}_0(t) = \Sigma_{i=1}^{n} \frac{I(T_i \leq t)\Delta_i}{\Sigma_{j \in R_i} e^{\hat{\beta} X_j}},$$

where T_i are ordered event times and R_i are observations still at risk.

3.2.2 Two-step estimation procedure

Note that, due to the form of most copula functions, the first and second derivatives of full log-likelihood function are typically quite complex. For example, the cumulative density function for Clayton copula (2.5) with respect to u and v is

$$c_\eta(u, v) = (1 + \eta)(u \cdot v)^{-1-\eta}(u^{-\eta} + v^{-\eta} - 1)^{-\frac{1}{\eta}-2}. \tag{3.3}$$

In the Appendix, we derive the first and second order derivatives of $C_\eta(u, v)$, which are very complex. This may cause issues in both estimating speed and stability. In addition, the analytical form is not reproducible when changing the copula function or the marginal distribution specification.

Then by formula (2.5) and (3.3), under Cox PH assumption, we can derive the explicit joint density function with covariates as

$$\begin{aligned} f(t_{1i}, t_{2i}|\beta, \eta, \lambda_0) &= (1 + \eta) \times \left[S_{01}(t_{1i})^{exp(\beta_{1i}x_{1i})} \times S_{02}(t_{2i})^{exp(\beta_{2i}x_{2i})}\right]^{-1-\eta} \\ &\times \left[S_{01}(t_{1i})^{-\eta \exp(\beta_{1i}x_{2i})} + S_{02}(t_{2i})^{-\eta \exp(\beta_{2i}x_{2i})} - 1\right]^{-\frac{1}{\eta}-2}. \end{aligned} \tag{3.4}$$

In order to derive the above score test statistic in (3.1), we need to estimate θ under H_0. Motivated by the two-stage estimation approach from Shih and Louis (1995), we propose a two-step maximum likelihood estimation procedure to obtain the restricted MLE $\hat{\theta}_0 =$

$(0, \hat{\theta}_{20})$. In step 1, we first obtain initial estimates of the parameters in marginal distributions (i.e., S_0 and β_{ng}) based on marginal likelihood functions. Then we maximize the pseudo joint likelihood (with the initial estimates of S_0 and β_{ng} plugged in) to get an initial estimate of the association parameter η. Then in step 2, we maximize the joint likelihood with the estimates from step 1 being initial values to obtain final estimate of $\hat{\theta}_0$. Detailed steps are provided below:

(1) Obtain initial estimates of θ_0:

a. $(\hat{\beta}_{ng}^{(1)}, \hat{\alpha}^{(1)}) = \underset{(\beta_{ng},\alpha)}{\arg\max} \log L_0(\beta_{ng}, \alpha)$, where L_0 denotes the likelihood function under marginal univariate data (without the genetic factor);

b. $\hat{\eta}^{(1)} = \underset{\eta}{\arg\max} \log L(\hat{\beta}_{ng}^{(1)}, \hat{\alpha}^{(1)}, \eta)$.

(2) Optimize the joint log-likelihood function of the bivariate data to get final estimates:
$\hat{\theta}_{20} = (\hat{\eta}, \hat{\alpha}, \hat{\beta}_{ng}) = \underset{(\beta_{ng},\alpha,\eta)}{\arg\max} \log L(\beta_{ng}, \alpha, \eta)$ with initial value $(\hat{\beta}_{ng}^{(1)}, \hat{\alpha}^{(1)}, \hat{\eta}^{(1)},)$.

Standard two-step estimation procedure for copula models stops after step (1-ii), since the association parameter η is of the primary interest. Note that, the initial estimates from step (1) are already consistent and asymptotically normal (Shih and Louis, 1995). But one cannot use Hessian matrices directly from step (1-i) and (1-ii) to obtain variance estimates for $(\hat{\alpha}, \hat{\beta}_{ng})$. The second step produces correct variance covariance estimates for all the parameters using the joint likelihood. Our experience finds that, the performance of convergence is greatly improved by using initial values from the first step.

For non-parametric baseline hazard function estimator, such as Breslow, one can obtain consistent estimates for regression coefficients and the association parameter by treating baseline hazards as nuisance parameters. However, the Hessian matrix from the PML in step 2 cannot be directly used for estimating the variance of $\hat{\beta}$ and $\hat{\eta}$. One solution is to use bootstrapped variance estimates, for example, see Lawless and Yilmaz (2011). Alternatively, a sieve-based smoothing technique can be used to estimate the baseline hazards (He and Lawless, 2003; Ding and Nan, 2011). In that case, the semiparametric M-estimation theory applies and the variance estimates for $\hat{\beta}$ and $\hat{\eta}$ can be obtained from the joint sieved log-likelihood in step 2.

3.2.3 Model selection

A crucial issue in applying the copula-based joint survival function to a given data set is how to choose a suitable parametric copula. Different copula functions can lead to very different dependence structures. Several model selection procedures have been proposed for copula-based time-to-event model. The Akaikes Information Criteria (AIC) (Akaike, 1998) and Bayesian Information Criteria (BIC) (Schwarz, 1978) have been widely used for model selection purpose in copula models. Wang and Wells (2000) proposed a model selection procedure based on nonparametric estimation of the bivariate joint survival function within Archimedean copulas. However, it requires to obtain a consistent estimate of a nonparametric bivariate joint survival function and its limiting distribution, which is often non-trivial in practice. Later, Chen et al. (2010b) proposed a penalized pseudo-likelihood ratio test with a two-step estimation approach for model selection under any parametric copula. To the best of our knowledge, there is no package exists for either of these two approaches, so we simply choose AIC statistics as our model selection criteria.

3.3 SIMULATION STUDIES

In this section, we evaluate the finite sample performance of the proposed test procedure through various simulation studies and compare it to the Wald test under Cox PH model with robust variance estimate (Lee et al., 1992). The Wald test from Cox model under independence assumption is also included for type-I error control simulations.

3.3.1 Bivariate time-to-event data generation

Recall that the bivariate joint survival function under a copula model is specified as

$$S(t_1, t_2) = C_\eta(S_1(t_1), S_2(t_2)),$$

where $U = S_1(T_1)$, $V = S_2(T_2)$ each follows a uniform distribution $U[0, 1]$. Define $W_v(u) = h(u, v) = P(U \leq u|V = v)$, which equals to $\partial C_\eta(u, v)/\partial v$. To generate bivariate survival data

(t_{1i}, t_{2i}), $i = 1, .., n$, we first generated v_i and w_i from two independent $U[0, 1]$ distributions. Then let $w_i = h(u_i, v_i)(= C_\eta(u_i, v_i)/\partial v_i)$ and solved for u_i from the inverse of h function h^{-1}. Finally, we obtained t_{1i} and t_{2i} from $S_1^{-1}(u_i)$ and $S_2^{-1}(v_i)$ respectively. We generated censoring times c_{1i} and c_{2i} from uniform distribution $U(0, C)$ with C chosen to yield censoring rates of 50%.

The value for the association parameter η was chosen to introduce weak or strong association, represented by Kendall's $\tau = 0.2$ and 0.6. We generated SNP data from a multinomial distribution with values $\{0, 1, 2\}$ and probabilities $\{p^2, 2p(1-p), (1-p)^2\}$, where p was the MAF, chosen to be 40% or 5%. We also included a continuous non-genetic risk factor $X_{ng,k}$ $(k = 1, 2)$, generated from a normal distribution $N(6, 2^2)$, where the mean and standard deviation were decided based upon our AMD data.

In all simulations, the sample size was $N = 500$ and we chose the same baseline marginal distribution for the two survival times (i.e., $S_{01}(t) = S_{02}(t)$). For type-I error control simulations, β_g was set to be 0. We replicated 100,000 runs and evaluated the type-I error at various α levels: 0.05, 10^{-2}, 10^{-3} and 10^{-4}, respectively. For power evaluation, we replicated 1000 runs under each SNP effect size. A range of β_g was picked to represent weak to strong SNP effect. For the situation when MAF = 0.4, β_g varies from 0.05 to 0.3. While for MAF = 0.05, β_g varies values from 0.05 to 0.6.

3.3.2 Parameter estimation

First, we examined the estimation statistics for baseline hazard coefficients $\alpha = (\lambda, k)$, copula association parameter η and genetic effect β_g. Table 1 reports the results from the situation where data were simulated from Clayton Weibull model. We fitted robust Cox, Clayton copula with Weibull, piecewise constant and Breslow margins models. Bootstrap variances were calculated for Breslow margins model. For baseline hazard parameters, Clayton Weibull model obtains good estimation accuracy at all settings. For β_g, the robust Cox, Clayton Weibull and Breslow margin models obtain more accurate estimates than the Clayton piecewise constant model. All four models achieve good coverage probabilities. For η, the Clayton Weibull model obtains more accurate estimates than the Clayton piecewise constant model.

Table 1: Estimation and inference statistics from Clayton copula models with different marginal distributions. True data were simulated from Clayton copula with Weibull marginal distributions.

	$\lambda = 0.1$				$k = 2$				$\eta = 0.5$				$\beta_{ng} = 0$				$\beta_g = 0$			
	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
Cox robust	-	-	-	-	-	-	-	-	-	-	-	-	0.0004	0.0241	0.0224	0.959	0.0015	0.0787	0.0740	0.938
Copula Weibull	0.0002	0.0075	0.0075	0.949	0.0092	0.0708	0.0722	0.956	0.0055	0.1237	0.1225	0.953	0.0005	0.0214	0.0213	0.960	0.0018	0.0763	0.0721	0.939
Copula PW	-	-	-	-	-	-	-	-	0.0390	0.1323	0.1313	0.947	0.0005	0.0208	0.0213	0.964	0.0024	0.0851	0.0830	0.955
Copula Breslow	-	-	-	-	-	-	-	-	0.0086	0.1239	0.1267	0.957	0.0004	0.0227	0.0222	0.945	0.0016	0.0744	0.0724	0.943
	$\lambda = 0.1$				$k = 2$				$\eta = 3$				$\beta_{ng} = 0$				$\beta_g = 0$			
	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
Cox robust	-	-	-	-	-	-	-	-	-	-	-	-	0.0006	0.0226	0.0224	0.952	0.0010	0.0936	0.0868	0.930
Copula Weibull	0.0001	0.0055	0.0055	0.955	0.0110	0.0784	0.0774	0.946	0.0172	0.333	0.336	0.951	0.0004	0.0120	0.0123	0.960	0.0022	0.0754	0.0710	0.932
Copula PW	-	-	-	-	-	-	-	-	0.0223	0.3583	0.3683	0.939	0.0004	0.0114	0.0120	0.962	0.0018	0.0733	0.0713	0.947
Copula Breslow	-	-	-	-	-	-	-	-	-0.0251	0.3586	0.3648	0.935	0.0003	0.0157	0.0166	0.958	0.0024	0.0851	0.0830	0.955
	$\lambda = 0.1$				$k = 2$				$\eta = 0.5$				$\beta_{ng} = 0$				$\beta_g = 0.5$			
	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
Cox robust	-	-	-	-	-	-	-	-	-	-	-	-	0.0011	0.0214	0.0207	0.942	0.0019	0.0757	0.0699	0.936
Copula Weibull	0.0003	0.0069	0.0070	0.9444	0.0072	0.0661	0.0664	0.937	0.0024	0.1113	0.1106	0.944	0.0013	0.0194	0.0193	0.949	0.0015	0.0715	0.0673	0.937
Copula PW	-	-	-	-	-	-	-	-	0.0457	0.0122	0.0120	0.945	0.0003	0.0186	0.0194	0.953	-0.0175	0.0697	0.0679	0.933
Copula Breslow	-	-	-	-	-	-	-	-	0.0041	0.1141	0.1125	0.946	0.0008	0.0208	0.0203	0.942	0.0014	0.0690	0.0735	0.934
	$\lambda = 0.1$				$k = 2$				$\eta = 3$				$\beta_{ng} = 0$				$\beta_g = 0.5$			
	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
Cox robust	-	-	-	-	-	-	-	-	-	-	-	-	0.0030	0.0209	0.0206	0.950	0.0050	0.0880	0.0816	0.936
Copula Weibull	0.0001	0.0050	0.0050	0.943	0.0092	0.0736	0.0719	0.938	0.0138	0.3020	0.3085	0.953	0.0021	0.0100	0.0100	0.952	0.0034	0.0687	0.0654	0.952
Copula PW	-	-	-	-	-	-	-	-	0.2265	0.3379	0.3431	0.924	0.0001	0.0097	0.0095	0.961	-0.0151	0.0672	0.0662	0.950
Copula Breslow	-	-	-	-	-	-	-	-	-0.0440	0.3357	0.3322	0.937	0.0028	0.0137	0.0130	0.964	0.0046	0.0806	0.0827	0.933

Number of replications = 1000, number of subjects = 500, censoring rate = 50%, MAF = 40%.

SE: standard deviation of the point estimate, SEE: mean of the standard error estimates, CP: 95% coverage probability.

3.3.3 Computing time

Table 2: Computing time required for testing 1000 variants using different methods.

Model	Time (second)	Non-convergence runs (rate)
Cox Robust	12.23	11/1000 (1.1%)
Copula score	654.90	0/1000 (0 %)
Copula LRT	1272.90	0/1000 (0 %)
Copula Wald	1762.96	0/1000 (0 %)

* Number of subject =500, censoring rate =50%, MAF = 5%, Kendall's $\tau = 0.6$

** A binary covariate $\sim$ Bernoulli(0.5) is included in the marginal regression model

We performed a simulation study to get an estimate of the computing time required for each method. True data were simulated from the Clayton copula with the Weibull margins model. All derivatives from the copula model were numerically solved. Table 2 shows that the Cox robust method is most computationally favorable. However, the convergence is an issue for the Cox robust model as we can see from the table. The non-convergence rate for the Cox robust model is 1.1%, which can be problematic when conducting a large number of tests on a genome-wide scale. Other limitations of the Cox robust method will be discussed in section 3.3.4. Within the copula framework, the score test used the least computing time compared to LRT and Wald test. This is because, for the the LRT and the the Wald test, we need to do parameter estimation under the alternative for each SNP. Specifically, for the Wald test, it also needs to calculate the Hessian matrix under each alternative. For the score test, although it needs to evaluate the score vector and the Hessian matrix for each SNP numerically, it only requires to perform the parameter estimation under the null once.

3.3.4 Simulation I: correctly specified model

In this section, we evaluated the method performance under correctly specified models. The true models are from the Clayton copula with the Weibull or Gompertz marginal distributions. With Weibull margin, we chose $\lambda = 0.01$ and $\gamma = 2$, and with the Gompertz margin,

we chose $\lambda = 0.2$ and $\gamma = 0.05$. In both scenarios, we also fitted marginal distributions with piecewise constant hazards.

Table 3 provides empirical type-I error rates under different α levels for four methods, namely, (1) the Cox model under independence assumption, (2) the Cox model with robust variance/covariance estimate, (3) the copula model with parametric marginal distributions (either Weibull or Gompertz), and (4) the copula model with piecewise constant marginal distributions. It is clearly seen that when MAF=40%, all the methods, except for the independent Cox method, control the type-I error well. However, when MAF=5%, the robust Cox method yields inflated type-I error rates at all α levels, especially when α level is low. For example, with data generated from Clayton copula with Weibull margins, the type-I error from the robust Cox method is 0.003 and 0.0007 for $\alpha = 0.001$ and 0.0001, respectively, which is 3 or 7 times of the expected value. The two copula methods control type-I error very well under both common and rare allele frequency scenarios. The independent Cox method always inflates the type-I error, which is not surprising.

Table 3: Type-I error for testing the genetic effect at various α levels for the Clayton copula with Weibull and Gompertz marginal distributions.

		Kendall's $\tau = 0.2$				Kendall's $\tau = 0.6$			
MAF	α	**Cox-I**	Cox-R	Cop-P	Cop-PW	Cox-I	Cox-R	Cop-P	Cop-PW
					Clayton Weibull				
5%	**0.05**	0.084	**0.062**	0.052	0.044	0.141	**0.063**	0.053	0.046
	0.01	0.022	**0.016**	0.011	0.009	0.053	**0.017**	0.012	0.009
	0.001	0.0034	**0.0029**	0.0012	0.0009	0.0129	**0.0030**	0.0014	0.0010
	0.0001	0.0007	**0.0007**	0.0001	0.0001	0.0035	**0.0007**	0.0002	0.0002
40%	**0.05**	0.086	0.054	0.052	0.045	0.142	0.055	0.053	0.046
	0.01	0.023	0.012	0.011	0.009	0.054	0.012	0.011	0.009
	0.001	0.0040	0.0015	0.0013	0.0008	0.0132	0.0014	0.0012	0.0009
	0.0001	0.0007	0.0002	0.0001	0.0001	0.0033	0.0001	0.0001	0.00004
					Clayton Gompertz				
5%	**0.05**	0.083	**0.061**	0.053	0.044	0.138	**0.062**	0.053	0.044
	0.01	0.022	**0.016**	0.011	0.008	0.051	**0.016**	0.011	0.009
	0.001	0.0034	**0.0029**	0.0012	0.0009	0.0129	**0.0030**	0.0014	0.0010
	0.0001	0.0006	**0.0006**	0.0002	0.0001	0.0032	**0.000**7	0.0002	0.0001
40%	**0.05**	0.084	0.054	0.052	0.044	0.140	0.054	0.052	0.044
	0.01	0.023	0.012	0.011	0.008	0.053	0.012	0.011	0.008
	0.001	0.0040	0.0015	0.0013	0.0008	0.0132	0.0014	0.0012	0.0009
	0.0001	0.0007	0.0002	0.0001	0.0001	0.0036	0.0001	0.0002	0.0001

* Non-genetic effect includes a margin-specific continuous variable and a shared cluster-specific variable.

** Number of replications = 100,000. Sample size = 500.

*** Cox-I (Independent Cox), Cox-R (Robust Cox), Cop-P (copula with a parametric margin) Cop-PW (coupla with a piecewise constant hazards margin).

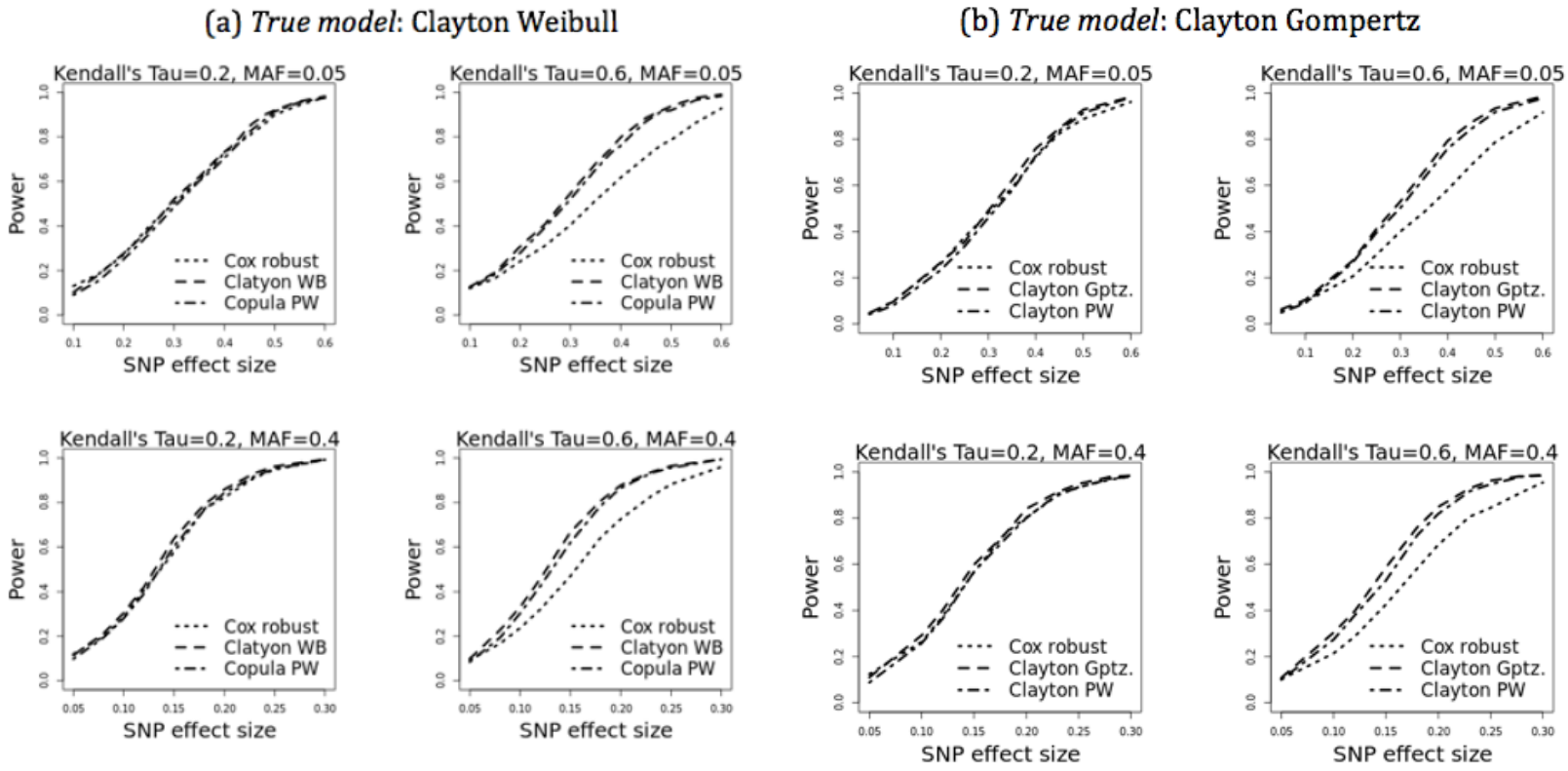

Figure 2: Simulation results for power comparison between robust Cox (Cox-R) model, copula models with parametric and weakly parametric (i.e., piecewise constant) margins over different genetic effect sizes with 5% α level.

Figure 2 presents the power curves over different genetic effect sizes for the three methods: robust Cox, Clayton copula with Weibull margins, and Clayton copula with piecewise constant margins. True models are under the Clayton copula with Weibull or Gompertz margins (Replicates = 1000. Sample size = 500, censoring rate = 50%). We can see that both copula methods yield better power as compared to the robust Cox method, especially when the association is strong. The parametric copula method is slightly more powerful than the weakly parametric copula model, which is as expected.

We also fitted the robust Weibull method for the case where the marginal distributions are Weibull. The results (in terms of both type I error control and power) are very close to the results from the robust Cox method and thus are omitted. Therefore, the inflated type-I error issue when MAF is small exists in the robust parametric (marginal) method as well.

3.3.5 Simulation II, misspecified model

In this section, we evaluated the method performance in the situation of misspecification of either the copula function or the marginal distributions. In the case of the copula function being misspecified, data were generated from a Gumbel copula with Weibull marginal distributions where $\lambda = 0.01$ and $\gamma = 2$. For misspecification of the marginal distribution, data were generated from a the Clayton copula with Gompertz distributions where $\lambda = 0.2$ and $\gamma = 0.05$. In both scenarios, data were fitted by Clayton copula with Weibull marginal distributions or piecewise constant marginal distributions.

Table 4 provides type-I errors under different α levels for two misspecified scenarios. The same four methods as in section 3.3.4 were compared. Under both scenarios, the two Cox model approaches do not depend on copula model specifications (so long as the marginal distributions are still from the PH family), and thus yield similar performance as those under correct model specification cases. When copula function is misspecified, the parametric copula model shows an obvious inflation on type-I errors, especially when the association is strong. Copula model with piecewise constant margins shows a smaller degree of inflation of type-I error rates. When marginal distributions were misspecified, the parametric copula shows a conservative type-I error control. The copula model with piecewise constant mar-

gins is robust against incorrectly specified marginal distributions. Therefore, our proposed method depends on correctly specified copula model, and thus, model selection (diagnostics) is recommended in real data analysis.

3.4 APPLICATION ON AREDS DATA TO IDENTIFY RISK VARIANTS FOR AMD PROGRESSION

3.4.1 AREDS data analysis

We applied our methods to the Age-related Eye Disease Study (AREDS) data. AREDS was a multi-center, controlled, randomized clinical trial sponsored by National Eye Institute (Age-Related Eye Disease Study Research Group, 1999). It was designed to assess the clinical course of, and risk factors for the development and progression of AMD. Participants in this study were followed up to 12 years, with scheduled visits every 6 months. In our research, we are interested in identifying genetic variants that are associated with progression to late AMD. It has been known that the progression profiles of two eyes are correlated.

The study collected DNA samples of consenting participants centrally by the International AMD Genomics Consortium (Fritsche et al., 2016) and performed genome-wide genotyping. The Illumina platform with a custom-modified HumanCoreExome array was used to obtain the genotypes. Then imputation was performed using the 1000 Genomes Project reference panel (Phase I). Traditional genotypes data are set to values of 0: no minor allele; 1: one copy of the minor allele; 2: two copies of the minor allele. Here we use dosage data that is continuous and can be any number between 0 and 2. The final data set includes 8,974,355 SNPs (265,096 genotyped and 8,709,259 imputed SNPs).

We analyzed a subset of 629 Caucasian participants who have at least one eye in moderate AMD stage at baseline. For the purposed of saving computing time, we use this subset of data that is considered more at risk for analysis in this section. The time-to-progression was calculated for each eye of these participants. The overall censoring rate was 54% for our analysis sample. In this work, we specifically tested the common variants (i.e. SNPs

Table 4: Type-I error at various α levels with misspecified copula models. Data are generated from a) Clayton copula with Gompertz margin or b) Gumbel copula with Weibull margin.

		Kendall's $\tau = 0.2$				Kendall's $\tau = 0.6$			
MAF	Tail	**Cox-I**	Cox-R	Cop-P	Cop-PW	Cox-I	Cox-R	Cop-P	Cop-PW
					Misspecification of margin				
5%	**0.05**	0.085	0.063	0.032	0.043	0.142	**0.063**	**0.032**	0.043
	0.01	0.023	0.016	0.005	0.008	0.053	**0.017**	**0.005**	0.009
	0.001	0.0035	0.0030	0.0003	0.0007	0.0133	0.0034	0.0014	0.0008
	0.0001	0.0006	0.0006	< 0.0001	0.0001	0.0035	0.0007	0.0002	0.0001
40%	**0.05**	0.085	0.054	0.031	0.042	0.142	0.054	0.053	0.043
	0.01	0.024	0.012	0.005	0.008	0.053	0.012	0.005	0.009
	0.001	0.0041	0.0015	0.003	0.0007	0.0142	0.0016	**0.0004**	0.0008
	0.0001	0.0008	0.0002	< 0.0001	0.0001	0.0037	0.0002	< 0.0001	0.0001
					Misspecification of copula				
5%	**0.05**	0.079	0.060	0.058	0.049	0.134	0.060	**0.096**	**0.067**
	0.01	0.021	0.015	0.014	0.011	0.049	0.015	**0.030**	0.018
	0.001	0.0035	0.0030	0.003	0.0007	0.0133	0.0034	**0.0143**	0.0008
	0.0001	0.0007	0.0005	0.0003	0.0003	0.0035	0.0005	**0.0017**	**0.0007**
40%	**0.05**	0.077	0.052	0.056	0.048	0.133	0.052	**0.092**	**0.064**
	0.01	0.020	0.011	0.012	0.009	0.048	0.011	**0.027**	0.015
	0.001	0.0041	0.0015	0.0031	0.0007	0.0142	0.0016	**0.0042**	**0.0071**
	0.0001	0.0005	0.0001	0.0009	0.0001	0.0025	0.0001	**0.0009**	0.0002

* Clayton copula with Weibull margin was fitted as the Copula parametric (Cop-P) model in both scenarios.

** Clayton copula with piecewise constant hazards margin (Cop-PW) was also fitted in both scenarios.

with MAF $\geq$ 5%) from chromosome 1 and 10, since some of the most significant regions associated with AMD risk (i.e., the CFH and $ARMS2$ gene region) is on chromosome 1 and 10. In total, we analyzed around 840,000 SNPs.

To decide which non-genetic risk factors to include in the regression model, we performed univariate analysis using both Clayton copula with a Weibull margin model and the robust Cox model (Table 5). Variables with a univariate pvalue < 0.05 were considered as covariates in the multivariate copula model. The baseline severity score, age and the current smoker category had significant univariate p-values from both copula and robust Cox models. Although smoking has been known as a major risk factor for AMD, in this analysis, we did not include smoking in the multivariate model, for two reasons: 1) the current smoker group only accounted for 5% of the total sample, 2) computing time will be greatly increased with two more covariates. The treatment effects were adjusted by baseline AMD severity score to accommodate the stratified randomization (participants with less severe AMD at baseline were only randomized to placebo or antioxidants alone arms) (Age-Related Eye Disease Study Research Group, 1999). Therefore, the non-genetic risk factors we included are baseline age and baseline AMD severity score. The AMD severity was calculated based on centralized grading of stereoscopic fundus photographs. The severity score ranges from 1 to 12, with 12 being the most severe stage. Late AMD is defined as the severity score $\geq$ 9 (9: noncentral Geographic Atrophy (GA), 10: central GA, 11: Choroidal neovascularization (CNV), and 12: CNV and central GA). For each eye that is free of late AMD at baseline, the progression time is defined as the time (in years) from the baseline visit to the first visit that the severity score reaches 9 or higher. If the eyes severity score does not exceed 9 during the follow up, it is treated as censored with censored time defined as the last visit time.

To decide which copula function and which marginal distribution to choose for this dataset, we used AIC, given by $AIC = -2 \log L(D; \hat{\theta}) + 2k$, where k is the total number of parameters in θ_0. Specifically, we considered two copula functions, Clayton and Gumbel, and three marginal distributions: Weibull, Gompertz and piecewise constant.

Table 5: Univariate analysis for risk factors of progression-to-late-AMD using the Clayton copula model with Weibull margins.

Variable	Mean(SD)/N(%)	HR (95% CI)**	p (copula)	p (CoxRst)
Baseline age (year)	69.55 (5.23)	1.03 (1.01, 1.05)	2.6×10^{-3}	2.9×10^{-4}
Sex				
Male	269 (43%)	Reference		
Female	360 (57%)	1.20 (0.97, 1.44)	0.07	0.16
Baseline smoking				
Never	272 (43%)	Reference		
Former	324 (52%)	1.15 (0.96, 1.32)	0.13	0.14
Current	33 (5%)	1.86 (1.32, 2.62)	3.5×10^{-4}	1.6×10^{-3}
Education				
$\leq$ high school	223 (35%)	Reference		
$>$ high school	406 (65%)	0.85 (0.71, 1.01)	0.06	0.08
Baseline severity score	5.81 (1.27)	1.59 (1.46, 1.73)	3.3×10^{-25}	4.6×10^{-50}
Treatment*				
Placebo	149 (24%)	Reference		
Antioxidants only	159 (25%)	0.81(0.64, 1.03)	0.09	0.37
Zinc only	157(25%)	1.13 (0.89, 1.45)	0.31	0.77
Antioxidants + zinc	164 (26%)	0.98 (0.77, 1.24)	0.85	0.99

*Treatment effect is adjusted by baseline AMD severity score

**Hazard ratio is computed based on a Clayton copula with a Weibull margin

*** CoxRst: Cox robust model with adjusted variance estimates

3.4.2 Data analysis results

Table 6 presents the AIC values for each model under the null hypothesis with non-genetic risk factors only. The Weibull marginal distribution under both copula models produced similar AIC values, which are smaller than other AIC values. We performed analyses using both Gumbel and Clayton copulas and their results are very similar. We presented the results from Clayton copula (with Weibull margin), given the computation time for the Clayton copula is much faster than for the Gumbel copula.

Table 6: The AIC values for the candidate models under null hypothesis with non-genetic risk factors only (i.e., baseline age and baseline severity scores).

Marginal Dist.	**Clayton copula**	**Gumbel copula**
Weibull	4442.524	4441.449
Gompertz	4485.848	4463.890
Piecewise Constant	4544.953	4508.787

Table 7 presents the 10 top significant (independent) variants discovered from our analysis or known variants associated with AMD disease risk. The variant $rs10922109$ is a known common variant associated with AMD disease risk from the CFH region with MAF = 28%. Note that, the known variant $rs10922109$ from CFH on chromosome 1 does not rank within the top 10 in this sub-population GWAS result. It still has a small p-value of 3.3×10^{-4} with our proposed method. The variant $rs2672599$ is another known common variant associated with AMD disease risk from the ARMS2/HTRA1 region with MAF = 35% respectively. For SNP $rs2672599$ the estimated effect size (i.e., Hazard Ratio) for this SNP is 1.42, with a 95% CI = $[1.23, 1.65]$. Fig. 4(a) is the marginal (eye-level) Kaplan-Meier (K-M) plot, which shows this variant can separate AMD progression curves quite well. The p-values from the copula-based method are slightly smaller than those from the robust Cox model for most of those 10 SNPs.

Figure 3 shows Manhattan plots for all variants with MAF $\geq$ 5% on chromosome 10 tested by robust Cox model and Copula Weibull model respectively. The Manhattan plots

Table 7: The p-values from robust Cox and Clayton copula with Weibull margins for the 10 top SNPs on chromosome 1 and 10.

SNP	Gene	BP	MAF	CoxRst	ClaytonWeibull
		CHROMOSOME 1			
$rs12083705$	KMO	241,715,015	0.09	7.1×10^{-7}	2.0×10^{-5}
$rs74960672$	$PLXNA2$	209,019,110	0.05	7.8×10^{-6}	1.6×10^{-6}
$rs5003371$	$CHI3L2$	111,784,260	0.08	7.8×10^{-5}	2.7×10^{-6}
$rs2206514$	$PTCHD2$	11,642,294	0.33	6.4×10^{-5}	4.5×10^{-5}
$rs12757114$	$LOC10012913$	105,670,026	0.47	2.4×10^{-5}	8.5×10^{-7}
		CHROMOSOME 10			
$rs72798393$	$LOC101928913$	67,031,293	0.09	3.3×10^{-5}	5.5×10^{-7}
$rs73292512$	$C10orf11$	78,171,701	0.05	2.4×10^{-5}	8.5×10^{-7}
$\boldsymbol{rs2672599}$	$\boldsymbol{ARMS2/HTRA1}$	124,211,625	0.35	2.1×10^{-5}	2.7×10^{-6}
$\boldsymbol{rs2284665}$	$\boldsymbol{HTRA1}$	124,226,380	0.33	8.4×10^{-5}	3.0×10^{-6}
$rs10828143$	$SLC39A12$	18,338,012	0.15	4.6×10^{-5}	5.1×10^{-5}

* Bolded regions are known regions associated with AMD prevalence

suggest that our proposed copula method obtains more signals compared to Cox model. In either method, none of the SNPs achieves a p-value $< 10^{-8}$, which is a commonly accepted GWAS significance level. Using the suggestive GWAS level of 10^{-5}, on chromosome 10, there are 58 variants with $p < 10^{-5}$ identified from the copula-based approach, while no variants with $p < 10^{-5}$ identified from Cox robust method on chromosome 10. On chromosome 1, neither methods has identified variant with p-value $< 10^{-8}$, while there are 15 variants with p-value $< 10^{-5}$ using the Clayton copula with Weibull margin and 7 variants when using the robust Cox model.

In addition to the test results for each variant, we can obtain both estimated joint and conditional survival functions from the copula model with Wald test on variants of interest, which can be used to establish a predictive model for progression-free probabilities. For example, Fig. 4(b) plots the joint 5-year progression-free probability contours (i.e, neither eye is progressed by year 5) for subjects having the same baseline severity score (=5.8) and age (=69.6) but different genotypes of the variant $rs2672599$. Fig. 4(c) plots the conditional 5-year progression-free probability of the remaining years for one eye, given the other eye has been progressed at year 5. It is clearly seen that in both plots, the three genotype groups are well separated, with the AA group having the largest progression-free probabilities.

We further picked two variants, $rs72798393$ from gene $LOC101928913$ and $rs2672599$ from gene $ARMS2$, and plotted the predicted 5-year joint progression-free probabilities by genotype group, varying the eye-level baseline severity score values (Fig. 5). We can see that carrying more T allele of $rs72798393$ leads to larger progression-free probabilities, indicated by the overall lighter color of the plot, while carrying more C allele of $rs2672599$ leads to smaller progression-free probabilities, indicated by the overall darker color of the plot. Within each genotype group, having a larger baseline severity score leads to smaller progression-free probabilities.

Moreover, in Fig.6, we plotted the predicted joint progression-free probability function $P(t_{1,i-1} < t_1 < t_{1,i}, t_{2,i-1} < t_2 < t_{2,i})$ by varying the values for $(t_{1,i-1}, t_{1,i}, t_{2,i-1}, t_{2,i})$ for subjects in different genotype groups of $rs2672599$. It is clearly seen that the joint progression-free probabilities decrease as the years increase, with smaller overall probabilities in subjects carrying more C alleles. We can also see that the two eyes are more likely to progress within

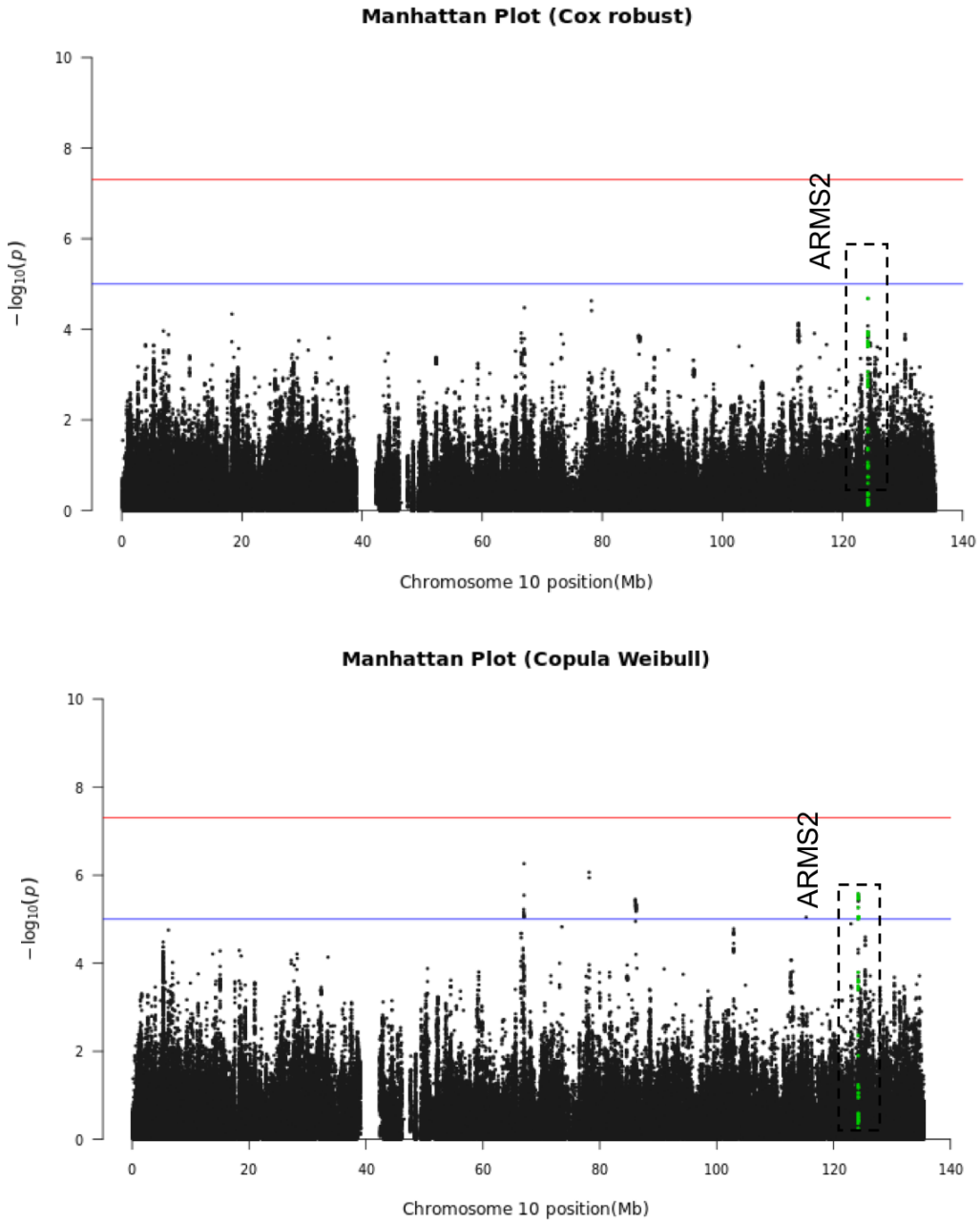

Figure 3: Manhattan plots of $-\log_{10}$(p-value) for all common variants (MAF $> 5\%$) on chromosome 10 from the AREDS data.

the similarly years, observed by the darker color cloud around the diagonal lines, which indicates the two eyes are correlated in terms of progression.

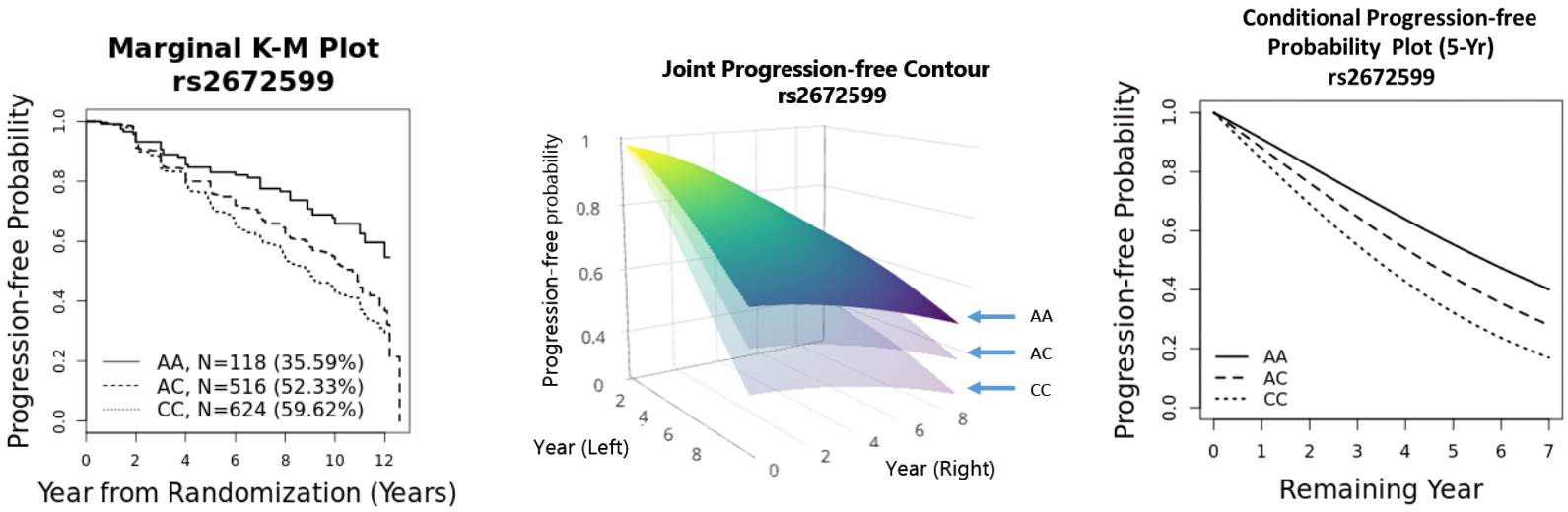

Figure 4: The estimated AMD progression profiles by a top SNP *rs*2672599 (*ARMS2*)

(a) The eye-level K-M plot, with the total number of eyes and the percent of progressed eyes in each genetic group given; (b) The copula-based estimated joint progression-free probability contours (the baseline severity score and age are fixed at their mean values: 5.8 and 69.6, respectively); (c) The estimated conditional progression-free probabilities of remaining years (since year 5) for one eye, given the other eye has been progressed by 5 year (the baseline severity score and age are also fixed at their mean values: 5.8 and 69.6, respectively).

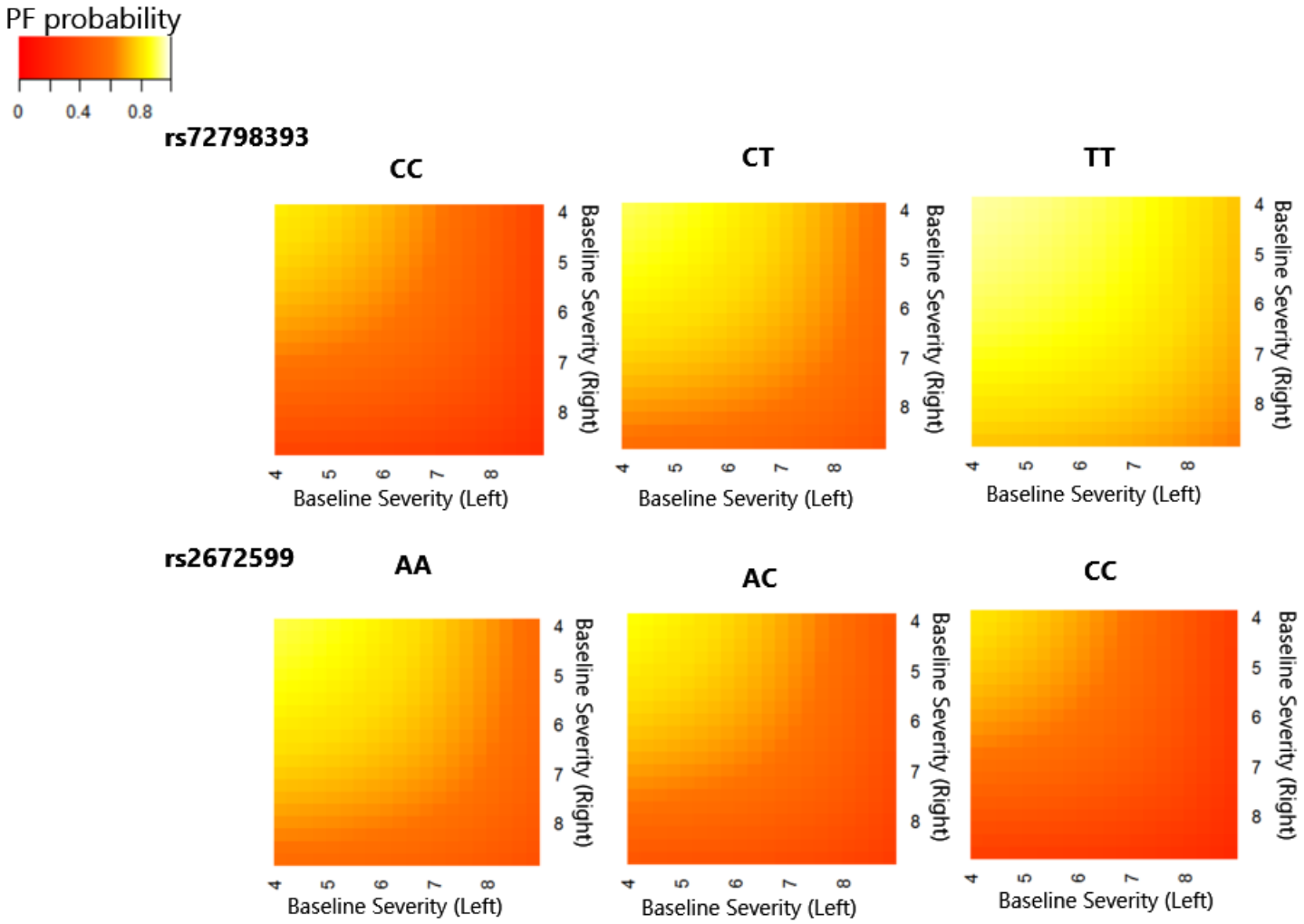

Figure 5: Predicted joint 5-year progression-free probabilities $P(T_1 > 5, T_2 > 5)$ for subjects with mean age 70 and baseline severity scores between 4 and 8 for both eyes, separated by genotype group of $rs72798393$ (gene: $LOC101928913$) and $rs2672599$ (gene: $ARMS2$), respectively.

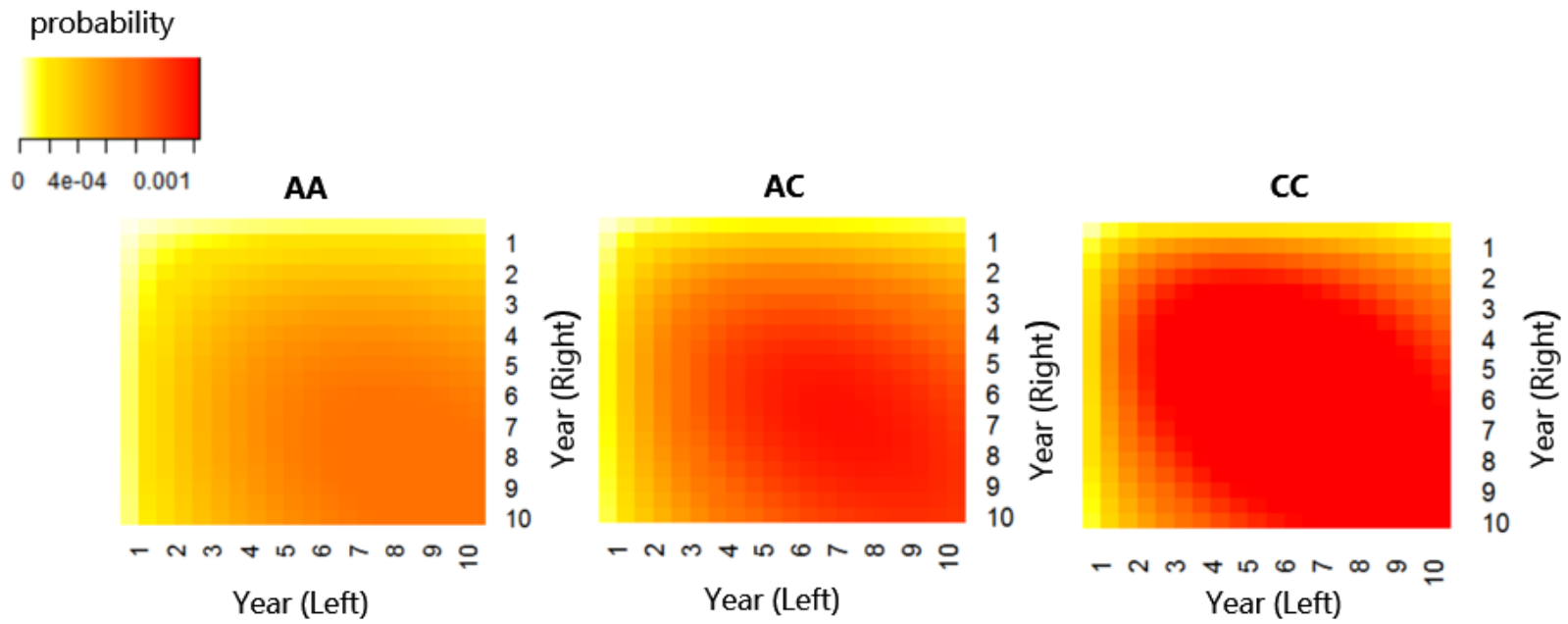

Figure 6: Predicted joint progression-free probabilities $P(t_{1,i-1} < t_1 < t_{1,i}, t_{2,i-1} < t_2 < t_{2,i})$ for subjects in different genotype groups of $rs2672599$ (gene: $ARMS2$). The baseline severity score and age are fixed at their mean values: 5.8 and 69.6, respectively.

Finally, to evaluate the model fitting, we plotted the estimated baseline hazard function from the copula model and displayed it on top of the K-M estimated baseline hazard function. Figure 7 shows that the two curves agree well with each other, indicating the Weibull margin with Clayton copula fits the data well. The estimated association parameter η is 1.23, which corresponds to Kendall's $\tau = 0.38$. This implies that there exists moderate association between the progression time of the two eyes.

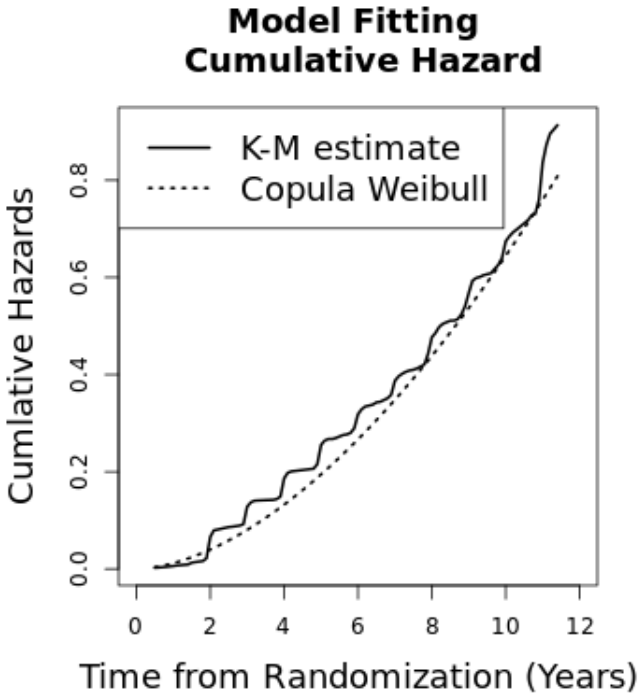

Figure 7: Estimated baseline hazard function from the Clayton copula with Weibull margins model and the empirical K-M hazard function estimates

3.5 DISCUSSION

Here we developed a computationally efficient copula-based score test procedure for large-scale bivariate time-to-event data. The copula model provides flexibility in modeling the association and marginal distributions separately. The score test procedure, as compared

to the other likelihood-based tests, has significant benefit in computation speed for the GWAS setting that we consider here. The proposed method has been demonstrated to be stable, with correct type-I error control and satisfactory power performance when the model assumptions are met.

Compared to the Cox model with robust variance estimates, which is frequently used in analyzing multivariate survival data, our copula-based method is more powerful when model is correctly specified. Moreover, our method is more suitable for testing low MAF variants as compared to robust Cox model in terms of type-I error control. Lastly, in our real data analysis, the robust Cox approach failed to converge in some situations when the MAF is low, while our copula-based approach was stable and unaffected by MAF.

Several directions may be pursued to extend the current proposed approach. First, as we have observed from our simulations, the performance of the proposed method is not robust against the misspecification of the copula function. Instead of using one-parameter copula functions as we consider here, one may consider using a two-parameter copula function, which is more flexible for characterizing the dependence structure of the bivariate data. For example, Chen (2012) have introduced a framework for estimating two-parameter copula models. In that setting, the association is described by two parameters in the copula function, with one to characterize the upper tail dependence and the other one to characterize the lower tail dependence. Both Clayton and Gumbel copulas are two limiting scenarios of the two-parameter copula family.

Secondly, we can further relax the assumption on marginal distribution to allow non-parametric or semiparametric estimates for the baseline hazard functions. We have examined the estimation performance using Breslow estimator for the baseline hazards function. Besides, one may consider relax the proportional hazards assumption to accelerated model or proportional odds models.

Lastly, in our AREDS data, the actual time-to-late-AMD is interval censored, due to intermittent assessment times (which were pre-determined in the trial). We currently treat them as right censored data given the interval lengths are fairly small and similar for all subjects. However, it is worthwhile to extend this test procedure to handle bivariate interval-censored data. All these directions are currently under investigation.

The application of the proposed method on AREDS data jointly modeled the progression profiles in both eyes, which, to the best of our knowledge, has not been done in any of the previous studies on AMD progression. The findings provided new insights about the genetic causes on AMD progression, which will be critical to establish novel and reliable predictive models of AMD progression to accurately identify high-risk patients at early stage. Our proposed methods are applicable to general bilateral diseases and are particularly useful for performing tests across a large number of biomarkers.

4.0 GENE-BASED TESTS FOR BIVARIATE TIME-TO-EVENT DATA THROUGH FUNCTIONAL REGRESSION

In genome-wide association studies, single variant test (SNP level) is very useful in detecting possible signals across the whole genome. Typically, after top loci are detected in replication studies, regions around top variants will then be fine-mapped to further evaluate the disease loci. However, there are some limitations when applying single variant tests. First, with each SNP been tested individually, due to its small effect size, it may suffer from an issue of lacking power and multiple testing. Secondly, true causal SNPs may not be genotyped due to the technology or cost reason. Instead, a SNP that is close to the true causal variant is often captured. With partial linkage disequilibrium (LD), the observed effect size is likely to be smaller. Furthermore, most statistical approaches used for single variant test are focused on common variants and can be too liberal when minor allele frequency is low. For example, we have identified in Chapter 3 that Cox model with robust variance adjustment leads to an inflated type-I error when MAF is small. A threshold for MAF is commonly applied when performing GWAS on single variants. In our single marker test scenario in Chapter 3 we applied a filter of MAF > 0.05 in all analyses. To solve these problems, there has been increasing interests in developing gene-based tests in the genetic analysis, which can usually take the LD information within a region into account and are suitable for collapsing a set of variants with low MAF.

The statistical methods for gene-based association studies can be broadly classified as burden tests, kernel-based association tests and functional linear model approaches, which we have introduced in section 2.4.3. Such approaches have been extended to censored time-to-event outcomes (Lin et al., 2011; Chen et al., 2014; Fan et al., 2016). It has been shown that for quantitative, binary and censored traits, functional linear can be more powerful

than SKAT type or burden test in many scenarios (Fan et al., 2013, 2014, 2016). The idea of treating the effect of genetic variants as an unknown function of actual physical positions in functional linear model utilizes the LD information among close variants.

Motivated by our research question on identifying associations between genetic markers and AMD progression, in this chapter, we develop several test procedures using functional regression (FR) for gene-based association analysis of bivariate censored traits.

4.1 COPULA-BASED FUNCTIONAL REGRESSION

4.1.1 Model specification

In order to perform tests on gene-based association analysis for bivariate time-to-event data, an intuitive extension is to combine the copula test procedure from Chapter 3 with the Cox functional regression model for censored traits as proposed by Fan et al. (2016).

Assume n individuals with m variants being sequenced for a genomic region. Physical positions for each variant within a region are denoted as $0 \leq u_1 < ... < u_m$ (which are typically standardized into $[0, 1]$). Let $G_i = (g_i(u_1), ..., g_i(u_m))', g_i(u_j) \in (0, 1, 2), j = 1, ..., m$ denote the genotypic information for the m variants, and $(X_{1i}, X_{2i}) = ((x_{1i1}, ..., x_{1ip}); (x_{2i1}, ..., x_{2ip}))$ denote a $p \times 2$ matrix of fixed effect covariates for subject i.

The bivariate Cox functional regression model on hazard function can be written as

$$\lambda_{ki}(t) = \lambda_{k0}(t) \exp\left(X_{ki}'\beta + \int_0^1 G_i(u)\gamma(u)du\right), k = 1, 2,$$

where $\lambda_{k0}(s)$ is the baseline hazard function for kth margin, β is a $p \times 1$ vector of coefficients for non-genetic fixed effect covariates, and $\gamma(u)$ is the genetic effect function of $G_i(u)$ at the position u. We assume both $\gamma(u)$ and $G_i(u)$ are smooth functions.

Then for each margin, the corresponding marginal survival function under Cox functional regression model is

$$\begin{aligned} S_{ki}(t) &= \exp(-\Lambda_{ki}(t)) = \exp\left(-\int \lambda_{ki}(t)dt\right) \\ &= \exp\left(-\int \lambda_{k0}(t) \exp\left(X_{ki}'\beta + \int_0^1 G_i(u)\gamma(u)du\right) dt\right). \end{aligned} \tag{4.1}$$

4.1.2 The genetic variant function $G_i(u)$

If the genotype data are of good quality with low missing rate, we can simply utilize the observed genetic information to represent $G_i(u)$. For example, we can use the raw genotype data as proposed in (2.10),

$$\hat{G}_i(u) = G_i = (g_i(u_1), ..., g_i(u_m))'.$$

If genotype data have a fairly high missing rate, we may apply an ordinary linear square smoother to obtain a continuous realization for discrete G_i. Let

$$\phi(u) = (\phi_1(u), ..., \phi_{B_x}(u))', k = 1, ..., B_x,$$

be a series of basis functions (e.g, B-spline or Fourier spline basis). Denote by Φ the $m \times B_x$ matrix with elements $\phi_b(u_j)$. Then through a linear square smoother (Ramsay et al., 2009), we can write an estimate of GVF $\hat{G}_i(u)$ as:

$$\hat{G}_i(u) = (g_i(u_1)), ..., g_i(u_m))\Phi[\Phi'\Phi]^{-1}\phi(u). \tag{4.2}$$

If missingness occurs,

$$\hat{G}_i(u) = (g_i(u_1)), ..., g_i(u_{m'}))\tilde{\Phi}[\tilde{\Phi}'\tilde{\Phi}]^{-1}\phi(u),$$

where $(g_i(u_1)), ..., g_i(u_{m'}))$ are observed non-missing genotypes and $\tilde{\Phi}$ is the corresponding basis matrix evaluated at observed genotypes.

4.1.3 The genetic effect function $\gamma(u)$.

The GEF $\gamma(u)$ is an unknown smooth function with an arbitrary form that we need to estimate. To do this one can expand it into a linear combination of basis functions and coefficients. Define a series of B_γ basis function by $\psi(u) = (\psi_1(u), ...\psi_{B_\gamma}(u))'$ and a $B_\gamma \times 1$ vector $\gamma = (\gamma_1, ..., \gamma_{B_\gamma})'$, then $\gamma(u)$ can be approximated by

$$\hat{\gamma}(u) = (\psi_1(u), ...\psi_{B_\gamma}(u))(\tilde{\gamma}_1, ..., \tilde{\gamma}_{B_\gamma})'. \tag{4.3}$$

Similarly, choices of $\psi_i(u)$ can be B-spline or Fourier basis. A standard test procedure to test whether the variants in a region are associated with the outcome can be translated to the null hypothesis:

$$H_0 : \gamma_1 = ... = \gamma_{B_\gamma} = 0.$$

4.1.4 Functional regression for hazard function

Depending on whether or not to smooth the GVF within the hazard function, we propose two types of functional regression models. The first approach is to smooth both $G(u)$ and $\gamma(u)$. Replace $G(u)$ and $\gamma(u)$ by expression in (4.2) and (4.3), we can have the Cox PH model formulated as:

$$\begin{aligned} \lambda_{ki}(t) &= \lambda_{k0}(t) \exp\left(X_{ki}'\beta + (g_i(u_1), ..., g_i(u_m))\Phi[\Phi'\Phi]^{-1}\gamma \int_0^1 \phi(u)\psi(u)du\right) \\ &= \lambda_{k0}(t) \exp(X_i'\beta + M_i'\gamma), \end{aligned} \tag{4.4}$$

where $M_i' = (g_i(u_1), ..., g_i(u_m))\Phi[\Phi'\Phi]^{-1} \int_0^1 \phi(u)\psi(u)du$. The integral $\int_0^1 \phi(u)\psi(u)du$ can be readily calculated using R package *"fda"* (Ramsay et al., 2009).

Another approach is to smooth $\gamma(u)$ only. In this formulation, we use

$$G_i = (g_i(u_1), ..., g_i(u_m))'$$

as a discrete approximation of $G(u)$ and only replace GEF with (4.3):

$$\begin{aligned}\lambda_{ki}(t) = \quad & \lambda_{k0}(t)\exp\left(X_{ki}'\beta + \left[\sum_{j=1}^{m}(g_i(u_j)\times(\psi_1(u_j),...,\psi_{B_\gamma}(u_j))\right](\gamma_1,...,\gamma_{B_\gamma})'\right) \\ & = \lambda_{k0}(t)\exp(X_i'\beta + M_i'\gamma), \end{aligned} \tag{4.5}$$

where $M_i' = \sum_{j=1}^{m} g_i(u_j)\times(\psi_1(u_j),...,\psi_{B_\gamma}(u_j))$ is a fully observed term.

As mentioned before, two common basis functions are the B-spline and the Fourier spline. The B-spline basis (de Boor, 2011) is a series of non-periodic functions with polynomial segments joint at values called knots. The segments have specifiable smoothness across every breaks. Advantages of applying the B-spline basis are its fast computation feature and great flexibility in structure. The Fourier splines consist of a set of periodic functions, with the basis function $\Phi_1(u) = 1$, $\Phi_{2r}(u) = \cos(2\pi ru)$ and $\Phi_{2r+1}(u) = \sin(2\pi ru)$, $r = 1,...,(B_\gamma - 1)/2$ (de Boor, 2011).

4.1.5 Bivariate functional regression under copula framework

In this work, the obtained genotype data from AREDS have passed strict quality control procedures and with no missings. Thus we do not smooth the GVF and use the observed G_i value for $G_i(u)$. So we only smooth the GEF. Note that, instead of estimating m parameters for m variants, we have only B_γ ($\ll$ m in most cases) parameters to estimate.

Recall that the bivariate joint survival function under a copula model can be written as a function of two marginal survival functions,

$$S(t_1,t_2) = C_\eta(S_1(t_1),S_2(t_2)),\ t_1,t_2 \geq 0. \tag{4.6}$$

Combine (4.1) and (4.6), we can write out the bivariate survival function with functional regression on marginal survival functions as:

$$\begin{aligned} S_i(t_{1i},t_{2i}) = C_\eta\Bigg(&\exp\left(-\int\lambda_{10}(t)\exp\left(X_{1i}'\beta + \int_0^1 G_i\gamma(u)du\right)dt\right), \\ &\exp\left(-\int\lambda_{20}(t)\exp\left(X_{2i}'\beta + \int_0^1 G_i\gamma(u)du\right)dt\right)\Bigg). \end{aligned}$$

Joint likelihood function $\mathcal{L}$ can be written out in similar style as in (2.5)

4.1.6 Score test

Now define the new parameter set $\theta = (\gamma, \beta, \alpha, \eta)$ and $\theta_0 = (\gamma = 0, \beta, \alpha, \eta)$, where $\alpha = (S_{01}, S_{02})$ are parameters in marginal proportion hazards model. Denote by $\theta_1 = \gamma$ the parameters of interest and $\theta_2 = (\beta, \alpha, \eta)$ the nuisance parameters. To test for association between the m genetic variants and the survival trait, the null hypothesis can be expressed as

$$H_0 : \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{B_\gamma})' = \mathbf{0}. \tag{4.7}$$

Note that, unlike the single marker test in Chapter 3, we are now simultaneously test a vector of parameters.

The corresponding score function and Fisher's information can be written as

$$U(\hat{\theta}_0) = \frac{\partial}{\partial \theta} \log L(D; \theta)\Big|_{\theta=\hat{\theta}_0} = (U_1'(\hat{\theta}_0), U_2'(\hat{\theta}_0))' = (U_1'(\hat{\theta}_0), 0')',$$

where $\hat{\theta}_0$ is the maximum likelihood estimator of θ under the restriction $\theta_1 = \gamma = \mathbf{0}$ (a $B_\gamma > 1$ dimension vector of zeros), $U_l(\cdot) = \partial \log L / \partial \theta_l,\ l = 1, 2$, and

$$\mathcal{I}(\hat{\theta}_0) = -E\left[\frac{\partial^2}{\partial \theta^T \partial \theta} \log L(D; \theta)\right]\Big|_{\theta=\hat{\theta}_0} = \begin{bmatrix} \mathcal{I}_{11}(\hat{\theta}_0) & \mathcal{I}_{12}(\hat{\theta}_0) \\ \mathcal{I}_{21}(\hat{\theta}_0) & \mathcal{I}_{22}(\hat{\theta}_0) \end{bmatrix},$$

with $\mathcal{I}_{11}, \mathcal{I}_{12}, \mathcal{I}_{21}$ and $\mathcal{I}_{22}$ being partitions of the information matrix $\mathcal{I}$ by θ_1 and θ_2.

$$Q_s = U_1'(\hat{\theta}_0)\mathcal{I}^{11}(\hat{\theta}_0)U_1(\hat{\theta}_0),$$

where $U_1'(\hat{\theta}_0)$ is a $B_\gamma \times 1$ vector and $\mathcal{I}^{11} = (\mathcal{I}^{-1})_{11} = (\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21})^{-1}$ is a $B_\gamma \times B_\gamma$ matrix.

Similar numerical approximation techniques from Chapter 3 are used for this score test approach. We use observed information matrix $\mathcal{J}(\hat{\theta}_0)$, where $\mathcal{J}(\theta) = -\frac{\partial^2 \log L(\theta; D)}{\partial \theta' \partial \theta}$, to approximate the information matrix. Then apply Richardson's extrapolation (Richardson, 1911)

to approximate the score function and observed information matrix, denoted by $\tilde{U}$ and $\tilde{\mathcal{J}}$. Therefore, the score test statistic we propose is

$$\tilde{Q}_s = \tilde{U}^T(\hat{\theta}_0)\tilde{\mathcal{J}}^{-1}(\hat{\theta}_0)\tilde{U}(\hat{\theta}_0) = \tilde{U}_1^T(\hat{\theta}_0)\tilde{\mathcal{J}}^{11}(\hat{\theta}_0)\tilde{U}_1(\hat{\theta}_0), \tag{4.8}$$

which asymptotically follows a $\chi^2_{B_\gamma}$ distribution with B_γ degrees of freedom under the null. We reject H_0 at the 5% level if $P(\chi^2_{B_\gamma} > \tilde{Q}_s) < 0.05$.

4.1.7 Likelihood ratio test

Different from the single marker test case, where the computing efficiency is a key factor for deciding the test procedure, we have many less tests in gene-based scenario ($\sim 20K$ genes in the whole genome). Therefore, an alternative approach is to perform likelihood ratio test, in addition to the score test.

With full log-likelihood function $\mathcal{L}(\theta)$, the likelihood ratio test statistic can be written as:

$$Q_l = -2(\mathcal{L}(\hat{\theta}_0) - \mathcal{L}(\hat{\theta})), \tag{4.9}$$

where $\mathcal{L}(\hat{\theta})$ is the the unrestricted maximum likelihood value and $\mathcal{L}(\hat{\theta}_0)$ is the restricted maximum likelihood with $\theta_1 = \gamma = 0$. Q_l also follows a $\chi^2_{B_\gamma}$ distribution with degrees of freedom B_γ and we reject H_0 at the 5% level if $P(\chi^2_{B_\gamma} > Q_l) < 0.05$.

4.2 FUNCTIONAL REGRESSION WITH COX ROBUST MODEL

In Chapter 2, we have mentioned the marginal model is a popular approach for handling dependence. Adapting the idea of marginal model, we can construct a Wald type test with a robust variance-covariance estimator as proposed by Lee et al. (1992) under the functional linear model framework. The null hypothesis is the same as (4.7).

Define $U(\gamma)$ to be the partial score vector under the marginal Cox PH model, we can modify formula (2.2) and get the partial score function under functional linear model as:

$$\begin{aligned} U(\beta,\gamma) &= \sum_{k=1}^{2}\sum_{i=1}^{n} U_{ki}(\beta,\gamma) \\ &= \sum_{k=1}^{2}\sum_{i=1}^{n} \Delta_{ki}\left\{\begin{pmatrix} X_{ki} \\ G_i \end{pmatrix} - \frac{S_k^{(1)}(t_{ki};\beta,\gamma)}{S_k^{(0)}(t_{ki};\beta,\gamma)}\right\} \\ &= \sum_{k=1}^{2}\sum_{i=1}^{n} \Delta_{ki}\left\{\begin{pmatrix} X_{ki} \\ G_i \end{pmatrix} - \frac{\sum_{p=1}^{n} Y_{kp}(t_{ki})\exp(X_{kp}'\beta + (G_p\Psi(u))'\gamma)\begin{pmatrix} X_{ki} \\ G_i \end{pmatrix}}{\sum_{p=1}^{n} Y_{kp}(t_{ki})\exp(X_{kp}'\beta + (G_p\Psi(u))'\gamma)}\right\}, \end{aligned} \tag{4.10}$$

where $Y_{ki}(t) = I\{T_{ki} \geq t\}$, $\Psi(u)$ is a $m \times B_\gamma$ matrix with element $\psi_b(u_j)$ for the bth basis function evaluated at the jth variant. Define $(\hat{\beta},\hat{\gamma})$ to be the root of $U(\beta,\gamma) = 0$. If the marginal Cox regression model is correctly specified, $n^{-\frac{1}{2}}U(\gamma)$ is asymptotically normally distributed with mean zero. The variance covariance matrix can be estimated by the "sandwich estimator"

$$\hat{\Gamma}(\hat{\gamma}) = \hat{A}(\hat{\gamma})^{-1}\hat{B}(\hat{\gamma})(\hat{A}(\hat{\gamma})^{-1})',$$

where $\hat{A}(\hat{\gamma}) = -\frac{1}{2n}\sum_{k=1}^{2}\sum_{i=1}^{n}\partial U_{ki}(\hat{\gamma})/\partial\gamma|_{\gamma=\hat{\gamma}}$ and $\hat{B}(\gamma) = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n}U_{ki}^2(\gamma)|_{\gamma=\hat{\gamma}}$.

A Wald type of test can be performed using this adjusted variance covariance matrix $\hat{\Gamma}(\gamma)$

$$Q_w = \hat{\gamma}\hat{\Gamma}^{-1}(\hat{\gamma})\hat{\gamma}' \sim \chi^2_{B_\gamma}. \tag{4.11}$$

Lee et al. (1992) has showed that Q_w in (4.11) asymptotically follows a χ^2 distribution with degrees of freedom B_γ.

4.3 SIMULATION STUDIES

In this section, we performed simulation studies to evaluate the performance of the proposed copula-based functional regression model. In our simulations, variants with MAFs less equal to 0.05 are considered as rare. Two scenarios were considered: (1) some causal variants are rare and some are common; (2) all causal variants are rare. We report the results of copula-based functional regression score and LRT test, functional regression with robust variance adjustment, univariate functional regression model, SKAT and burden test (from R package *SeqMeta* (Voorman et al., 2016)).

4.3.1 Data generation

Genetic data were generated from European ancestry of 10,000 haplotypes covering 1Mb regions, simulated by Yun Li at the University of North Carolina, Chapel Hill. Calibrated coalescent model as programmed in COSI was used to generate the haplotypes with linkage disequilibrium (LD) information (Schaffner et al., 2005). With 10,000 haplotypes, first we decided a genetic region of length 6 Kb and 30 Kb for all and rare only variants scenarios. For both all variant and rare only scenarios, the regions contains around 20 variants. A random mating technical was then applied to generate genetic information for n subjects.

In our simulation, we simulated 100 genetic datasets with a population of size 1000. For each genotype set, K replicates of bivariate survival times were generated using the same copula conditional distribution technique from section 3.3.1. Such approach lead to a total $100 \times K$ genotype-phenotype datasets and is similar to the SKAT paper of (Chen et al., 2014), which made comparison between methods more valid.

In this simulation study, we included one non-genetic effect, which was normal distributed with mean 6 and standard error 2. Bivariate event times were generated from a Clayton Weibull model with scale parameter $\lambda = 0.1$, rate parameter $k = 2$. A uniform distribution $[0, c]$ was used to generate censored times, where c was chosen to yield a censoring rate of 50%.

4.3.2 Type-I error

In type-I error analysis, marginal association levels corresponding to Kendall's $\tau = 0.05$, 0.4, 0.8 were evaluated. The marginal regression model was generated from

$$S(t_{ki}) = \exp\{-(\lambda t_{ki})^k e^{X'_{ki}\beta}\},$$

where β is the coefficient (chosen to be 0) for the non-genetic covariate.

Table 8 presents type-I error results for testing both common and rare variants ($MAF > 1\%$) in a region. It shows that when two margins are close to independent, i.e. Kendall's $\tau = 0.05$, all approaches obtain accurate type-I error rates. With association level increases, type-I error for Cox model under the independent assumption inflates quickly. The Cox robust method has slightly inflated type-I error rates. The copula-based models achieve good type-I error control rates at all association levels, despite the choice of spline. Within copula framework the score test and LRT obtain similar performance. However, such inflation does not increase with the association level.

Table 9 presents the type-I error results for testing only rare variants (MAF $\in [1\%, 5\%]$) in a region. The overall performance for the rare variant only situation is similar to the all variants scenario. Robust Cox FLM model are observed to have an inflation in type-I error rates, while all copula-based tests control the type-I error well at all nominal levels. Therefore, we can conclude that our proposed copula based methods are more suitable for testing rare variants.

4.3.3 Empirical power

In power analysis, we generated data evaluating both homogeneous genetic effects, i.e., genes with effect of the same direction, and heterogeneity genetic effects, i.e., genes with effects of opposite directions. The marginal regression model was generated from

$$S(t_{ki}) = \exp\{-(\lambda t_{ki})^k e^{X'_{ki}\beta + G'_i\gamma}\},$$

where $\gamma = (\gamma_1, ..., \gamma_{B_\gamma})$ are the coefficients for causal variants, $G = (G_1, ..., G_s)$ are s causal variants. The effect size for each causal variant was chosen to be $c^{\frac{|log_{10}MAF|}{2}}$, where

Table 8: Type-I error at various association levels from Clayton copula with Weibull margins for both common and rare variants.

		Copula FR			Cox FR		SKAT	Burden
Kendall's τ	α level	Copula-score	Copula-LRT	**CoxFLM-ind**	CoxFLM-single	**CoxFLM-Rst**	(Single)	(Single)
$\tau = 0.05$	0.05	0.0547	0.0524	0.0673	0.0523	0.0626	0.0494	0.0492
	0.01	0.0114	0.0102	0.0152	0.0104	0.0150	0.0099	0.0105
	0.001	0.0014	0.0012	0.0020	0.0012	0.0022	0.0012	0.0010
	0.0001	0.00015	0.00010	0.00033	0.00018	0.00047	0.00009	0.00010
$\tau = 0.4$	0.05	0.0527	0.0522	0.2119	0.0515	0.0653	0.0504	0.0495
	0.01	0.0105	0.0103	0.0832	0.0104	0.0162	0.0099	0.0099
	0.001	0.0011	0.0011	0.0214	0.0012	0.0025	0.0011	0.0010
	0.0001	0.00015	0.00012	0.00521	0.00010	0.00047	0.00009	0.00012
$\tau = 0.8$	0.05	0.0534	0.0521	0.3293	0.0526	0.0628	0.0511	0.0500
	0.01	0.0110	0.0105	0.1644	0.0107	0.0159	0.0101	0.0097
	0.001	0.0011	0.0011	0.0594	0.0010	0.0021	0.0012	0.0011
	0.0001	0.00012	0.00014	0.02103	0.00010	0.00042	0.00013	0.00012

Simulation setting: number of subjects = 1000, MAF > 1%, censoring rate = 50%, number of basis = 5, replication = 100,000

Copula-score: Score test using copula-based Cox proportional hazards functional linear model

Copula-LRT: likelihood ratio test using copula-based Cox proportional hazards functional linear model

CoxFLM-ind: LRT using Cox proportional hazards functional linear model treating two margins as independent

CoxFLM-single: LRT using Cox proportional hazards functional linear model using collapsed univariate cluster level data

CoxFLM-Rst: Wald test using Cox proportional hazards functional linear model with robust variance covariance adjustment

SKAT (single): SKAT with bivariate margins collapsed into univariate cluster level data

Burden (single): burden test with bivariate margins collapsed into univariate cluster level data

$c = 0.4, 0.3, 0.25$ for scenarios of 10%, 20% and 30% of causal variants in a given region, respectively.

Methods we compared include: subject level LRT using Cox FLM, subject level tests using SKAT, and burden tests, eye level Cox FLM with robust variance adjustment (Wald test), Copula model using Cox FLM score test, Copula model using Cox FLM LRT.

Table 9: Type-I error at various association levels from the Clayton copula with Weibull margins for rare variants.

		Copula FR			Cox FR		SKAT	Burden
Kendall's τ	α level	Copula-score	Copula-LRT	**CoxFLM-ind**	CoxFLM-single	**CoxFLM-Rst**	(Single)	(Single)
$\tau = 0.05$	0.05	0.0539	0.0540	0.0690	0.0516	0.0677	0.0490	0.0500
	0.01	0.0115	0.0118	0.0160	0.0105	0.0175	0.0094	0.0099
	0.001	0.0011	0.0011	0.0021	0.0010	0.0035	0.0009	0.0010
	0.0001	0.00014	0.00011	0.00030	0.00010	0.00135	0.00010	0.00009
$\tau = 0.4$	0.05	0.0541	0.0537	0.2151	0.0521	0.0674	0.0499	0.0505
	0.01	0.0113	0.0110	0.0855	0.0110	0.0173	0.0099	0.0101
	0.001	0.0013	0.0012	0.0226	0.0011	0.0032	0.0008	0.0011
	0.0001	0.00012	0.00012	0.00628	0.00010	0.00063	0.00014	0.00011
$\tau = 0.8$	0.05	0.0534	0.0517	0.3305	0.0519	0.0646	0.0502	0.0502
	0.01	0.0110	0.0101	0.1671	0.0112	0.0151	0.0103	0.0104
	0.001	0.0011	0.0010	0.0609	0.0012	0.0022	0.00091	0.00010
	0.0001	0.00012	0.00006	0.02122	0.0001	0.00035	0.00010	0.00008

Simulation setting: number of subjects = 1000, $MAF \in [1\%, 5\%]$, censoring rate = 50%, number of basis = 5, replication = 100,000

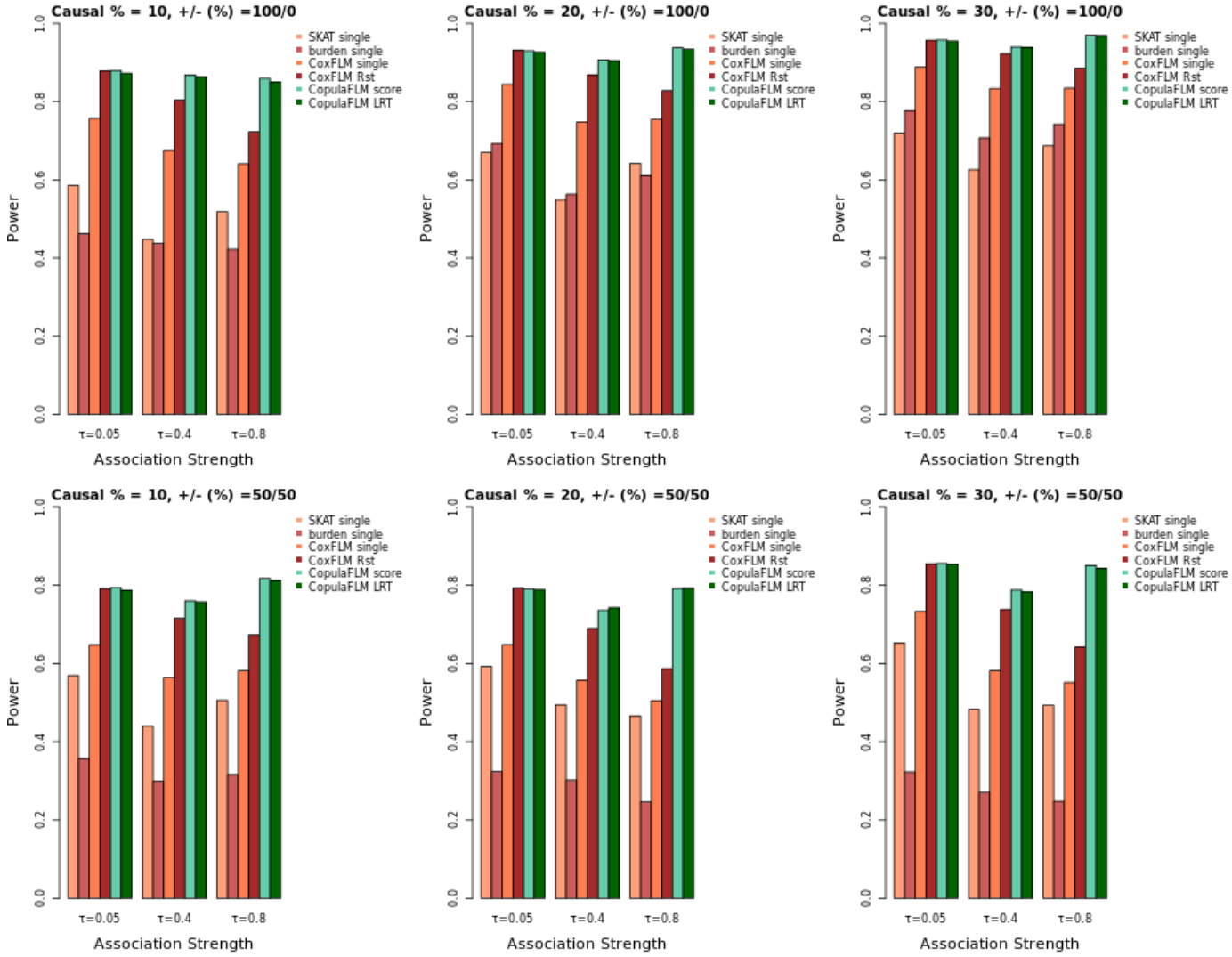

Figure 8: Empirical power analysis for 1000 gene regions at various association levels with both common and rare variants

Figure 8 presents the power bars over different association levels and genetic effect sizes for all variants ($MAF > 1\%$). Overall, we see bivariate methods achieve higher power than any univariate approaches. When association is strong, the copula-based test achieves better power than the robust Cox FLM method, indicating that the Cox robust method tended to be slightly conservative. We also note that, when there are heterogeneous genetic effects in a region, the power of burden test decreases significantly.

Figure 9 shows the power analysis over different association levels and genetic effect sizes using only rare variants ($1\% < MAF < 5\%$). We see that when variants are rare, within univariate methods, the Cox FLM is not superior to the SKAT or burden test. Especially, SKAT is very powerful for both cases of homogeneous and heterogeneous genetic effects. However, with bivariate approaches incorporated by either copula or robust method, power increases significantly and our proposed methods beat univariate SKAT/burden in most of cases.

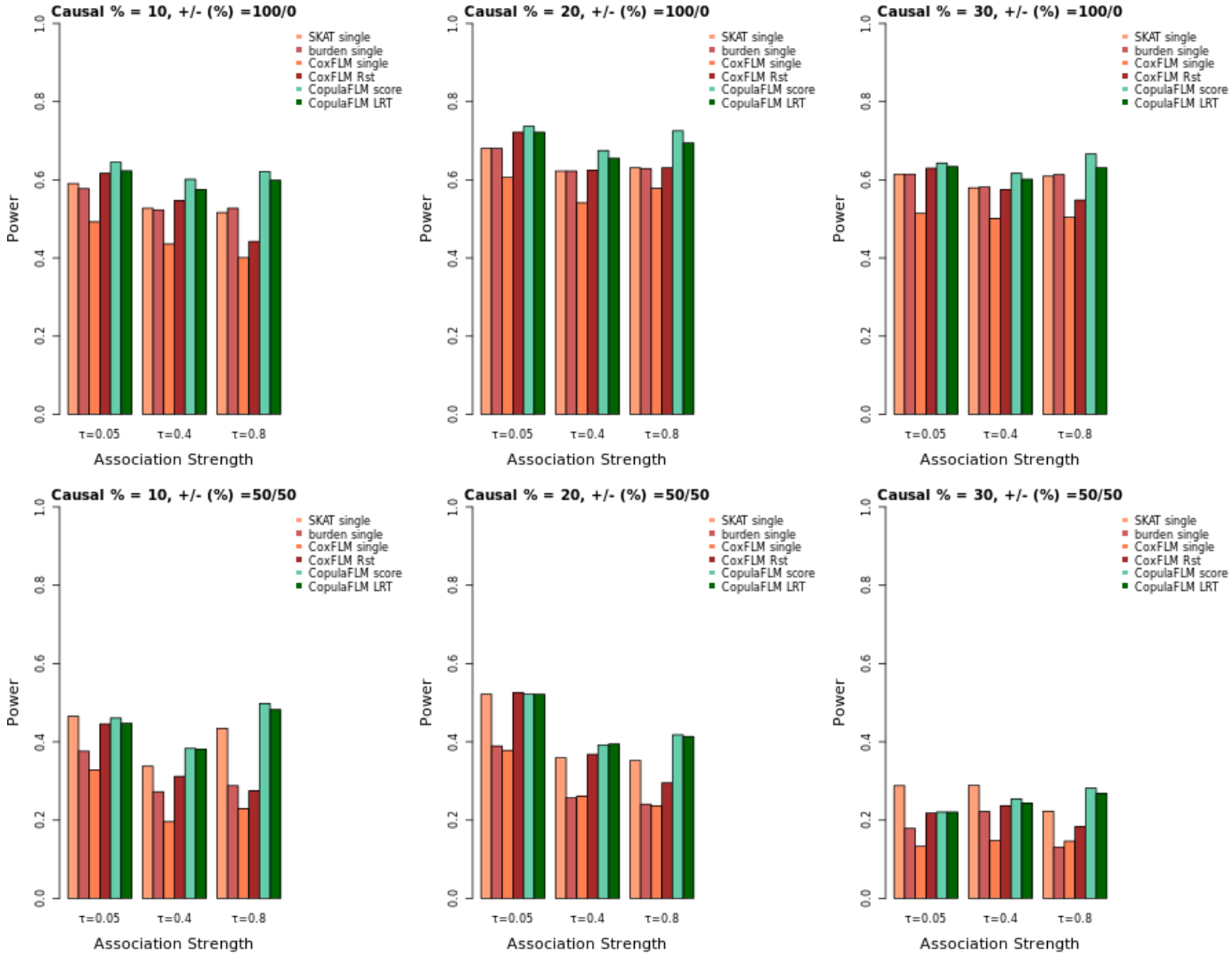

Figure 9: Empirical power analysis for 1000 gene regions at various association levels with rare variants only

4.4 REAL DATA ANALYSIS

4.4.1 AREDS data analysis

We applied our proposed methods on AREDS dataset (Age-Related Eye Disease Study Research Group, 1999) to identify gene regions that are associated with AMD progression. Detailed information about AREDS has been described in Chapter 3.

We included all Caucasian participants with neither eye progressed at the time of enrollment into the study. For bivariate approaches, time-to-progression was defined for each eye. For univariate approaches, we collapsed eye-level data into subejct level and calculated time to first eye progression as the progression time. A total of 2296 subjects were included in the analysis. The baseline age and severity score (on a continuous scale ranged from 1 to 8) were included in the regression part as non-genetic risk factors.

Table 10 presents top 4 genetic regions associated with AMD risk identified by single marker GWAS results in Fritsche et al. (2016). Therefore, we specifically examined these four regions with our methods. Location of gene regions are extracted based on GRCh37/hg19 assembly from UCSC genome broswer. Variants within $+/-5Kb$ of the region boundary were included in the analysis.

To be consistent with singler marker analysis in Chapter 3, we also fit a Clayton copula with Weibull margins in this gene-based analysis. B-spline functions were used to smooth the genetic effect function. The number of basis is usually decided by the total number of observations or through cross validation. In survival analysis it is related to the total number of events. In our analysis, we examined the scenarios with number of basis equals 5, 6 and 7. We compared our 3 methods: copula functional linear model based score test and likelihood ratio test, functional linear model with robust variance adjustment, with subject level univariate functional linear model from Fan et al. (2016), SKAT and burden test. For subject-level analysis, the phenotype data were defined as time to the first progressed eye.

Table 10: Top gene regions from single variant GWAS results

Region	Chr	StartPos**	EndPos**	Num all	Num rare***
*ARMS2 regions**	10	124,134,094	124,274,424	508	132
C2-CFB-SKIV2L	6	31,895,266	31,937,532	127	69
C3	19	6,677,846	6,720,662	296	49
CFH	1	196,621,008	196,716,634	261	52

* *PLEKHA1,MIR3941,ARMS2,HRTA1*

** Rare variant is defined by $1\% \leq \text{MAF} \leq 5\%$

*** Regions are selected with reference start and end pos $+/-5K$ (hg19)

4.4.2 Data analysis results

In both analyses with all variants and rare variants, we do not observe a specific pattern in p-values when varying the number of basis.

Table 11 presents result using all variants (including common and rare). Gene regions: CFH, C2 and ARMS2 are all significant with small p-values. Notice that the proposed bivariate approaches, i.e., copula score, copula likelihood ratio and Cox robust method achieve much smaller p-values than existing univariate approaches. For example, with 5 basis the p-values of copula FLM for ARMS2 regions are 1.1×10^{-4} and 1.0×10^{-4} with score test and LRT respectively. For Cox FLM with robust variance estimate, the p-value is slightly larger (6.4×10^{-3}). While for unvariate anslysis, Cox FLM achieves a p-value of 0.09.and neither SKAT or burden test reaches significance at α level of 0.05.

Table 12 presents the results using rare variants only to assess the genetic association with progression. We discover the p-values are larger than the cases with all variants. However, with only rare variants of MAF $\in [0.01, 0.05]$, our method can still identify most top regions as significant. The p-values from the bivariate copula-based methods and Cox robust FLM are also smaller than those univariate approaches.

Table 11: Bivariate functional regression results from AREDS data for top regions with both common and rare variants

		Copula FR		Cox FR		SKAT	Burden
Gene	basis number	CopFLM-score	CopFLM-lrt	CoxFLM-Rst	CoxFLM-single	(Single)	(Single)
CFH	5	1.4×10^{-9}	2.8×10^{-10}	6.4×10^{-7}	3.7×10^{-5}	0.03	2.2×10^{-3}
	6	3.5×10^{-9}	7.0×10^{-10}	1.5×10^{-6}	8.4×10^{-5}	0.03	2.2×10^{-3}
	7	1.1×10^{-8}	2.2×10^{-9}	4.1×10^{-6}	1.8×10^{-4}	0.03	2.2×10^{-3}
C3	5	0.06	0.06	0.04	1.8×10^{-3}	0.05	0.94
	6	0.12	0.13	0.08	4.1×10^{-3}	0.05	0.94
	7	0.10	0.10	0.06	2.1×10^{-3}	0.05	0.94
C2	5	1.6×10^{-5}	1.2×10^{-5}	3.9×10^{-4}	0.01	0.02	0.04
	6	1.6×10^{-5}	1.3×10^{-5}	3.5×10^{-4}	0.02	0.02	0.04
	7	3.3×10^{-5}	2.8×10^{-5}	7.0×10^{-4}	0.01	0.02	0.04
ARMS2	5	1.1×10^{-4}	1.0×10^{-4}	6.4×10^{-3}	0.09	0.10	0.31
	6	1.5×10^{-6}	1.8×10^{-6}	4.1×10^{-4}	1.8×10^{-3}	0.10	0.31
	7	5.4×10^{-4}	2.7×10^{-4}	9.9×10^{-3}	0.02	0.10	0.31

4.5 DISCUSSION

In this chapter, we extend the bivariate single variant test from Chapter 3 to gene-based test using the idea of the functional linear model. On one hand, under the functional linear model, the genetic effect can be viewed as a function of the physical positions of variants. On the other hand, the copula model can effectively handle the correlation between the margins. Combining FLM and the copula model will make it suitable for gene-based test on bivariate time-to-event data. In addition, we extend the univariate FLM test and derive a robust variance estimate (Liang and Zeger, 1986; Lee et al., 1992) for Wald test.

Extensive simulation studies were performed to evaluate the type-I error rates and power performance for our methods. Both the score test and the LRT tests in the context of copula FLM model control type-I error very well at all nominal levels. There is an noticeable inflation of type-I error for the Cox FLM with robust variance estimate. Such inflation

Table 12: Bivariate functional regression results from AREDS data for top regions with rare variants (MAF $\in [0.01, 0.05]$)

		Copula FR		Cox FR		SKAT	Burden
Gene	basis number	CopFLM-score	CopFLM-lrt	CoxFLM-Rst	CoxFLM-single	(Single)	(Single)
CFH	5	6.8×10^{-3}	4.5×10^{-3}	5.6×10^{-3}	0.10	3.0×10^{-3}	0.76
	6	8.4×10^{-3}	4.8×10^{-3}	8.7×10^{-3}	0.14	3.0×10^{-3}	0.76
	7	2.6×10^{-3}	9.1×10^{-4}	3.6×10^{-3}	0.11	3.0×10^{-3}	0.76
C3	5	0.53	0.50	0.57	0.83	0.13	0.31
	6	0.51	0.48	0.56	0.75	0.13	0.31
	7	0.45	0.42	0.62	0.90	0.13	0.31
C2	5	4.3×10^{-3}	2.2×10^{-3}	0.02	0.16	0.12	0.01
	6	8.7×10^{-3}	4.6×10^{-3}	0.03	0.23	0.12	0.01
	7	1.5×10^{-2}	8.3×10^{-3}	0.05	0.29	0.12	0.01
ARMS2	5	1.1×10^{-3}	1.1×10^{-3}	8.4×10^{-3}	0.03	0.23	0.11
	6	1.9×10^{-3}	1.4×10^{-3}	0.02	0.01	0.23	0.11
	7	5.4×10^{-4}	2.7×10^{-4}	9.9×10^{-3}	0.02	0.23	0.11

has also been observed in single variant scenario when the MAF is small in Table 3. For the power analysis, our bivariate tests show great advantage by utilizing all available data without collapsing them into subject level. Power is greatly increased in both scenarios that all causal variants are positive associated with phenotypes and causal variants have a mixture of positive/negative effects.

The great advantage of proposed copula/Cox FLM models is the genetic effects are treated as a function of the actual physical positions. Therefore, the LD information is accounted in this method. On the contrary, SKAT and burden test do not depend on any physical position. The test statistic for SKAT is a weighted sum of single marker score statistics (Chen et al., 2014), which only models the pairwise LD between markers. Fan et al. (2016) has demonstrated that FLM are more powerful than SKAT and the burden test in many univariate cases.

We also applied our methods and the other existing methods (i.e., subject-level burden test, SKAT, and FLM) on AREDS data to analyze genes in the whole genome to identify significant genes associated with AMD progression. Known AMD risk regions such as CFH and $ARMS2$ were identified as top genes from all methods. However, the genome-wide QQ-plots and the genetic control indices indicate inflated type-I errors for all methods including the subject-level SKAT and burden test. Whether this is due to some strange region sizes (i.e., the number of variants) or violation of certain model assumptions in this AREDS data needs further examination. Therefore, the GWAS results were not presented in this dissertation. We will investigate this gene-based GWAS on AREDS data and publish our findings in the manuscript based on contents from Chapter 4.

To the best of our knowledge, the bivariate gene-based test for censored data has not been done in any previous studies. Specifically, the findings from AREDS data provided new perspective about the genetic causes on AMD progression, which will be valuable to establish novel and reliable predictive models of AMD progression. The proposed methods can be used for genome-wide association study of any bilateral disease to identify disease susceptible genes.

5.0 CONCLUSION

The focus of this dissertation is to develop test procedures for genetic association analysis with bivariate time-to-event traits. In the first part, we proposed a computationally efficient score test for single variant that can be applied on genome-wide scale screening. In the second part, we proposed several gene-based test procedures under the framework of functional linear model. Our method can be applied to other bilateral diseases to identify genetic risk factors for time-to-event outcome.

We implemented our proposed test procedures for bivariate time-to-event data into a user friendly R package {CopulaTest}. The current version can handle both the score and Wald test under Clayton or Gumbel copulas with parametric, weakly parametric or non-parametric Cox PH margins. The gene-based test procedure will also be incorporated in the package, which will make this tool comprehensive to perform genetic association analysis for bivariate censored traits.

The work from Chapter 3 with copula-based score test for bivariate time-to-event data has been submitted to Journal of Royal Statistics Society: Series C and was under review at the time of dissertation submission. The manuscript is co-authored with Richard J Cook, (Department of Actuarial Science, University of Waterloo), Wei Chen (Department of Pediatrics, Children's Hospital of Pittsburgh) and Ying Ding (Department of Biostatistics, University of Pittsburgh).

5.1 FUTURE WORK

Several extensions to this work can be considered. Most of them are driven by features of our real data set.

First, the AREDS progression data are in fact interval-censored due to intermittent assessment times. We treated them as right-censored given the assessment intervals ($\sim$ 6 months) are relatively narrow compared to the entire follow-up time (12 years). However, it is worthwhile to extend the methods to handle interval censored data since such type of data often occur in practice. Thus extending our framework to make it suitable for interval censored data will be natural to consider in the next step.

Secondly, the specification of marginal distributions in copula model can be further relaxed. For example, one can consider the transformation model (Zhou et al., 2017) which includes both proportional hazards model and proportional odds model as special cases.

Thirdly, in this dissertation, we are not able to handle subjects with one eye already progressed at baseline. In real data analysis (AREDS), these subjects account for around 12% of the entire data. This data structure is a common phenomenon in bivariate censored data. A mixture model with a composite likelihood function can be considered to handle both two groups of subjects together: one with both margins and the other with only one margin.

Another interesting extension can be the use of a two-parameter copula. We have observed in Table 4 that misspecification on copula function can lead to serious problem in type-I error control. Sometimes one-parameter copula may not be sufficient to model the marginal dependence. With two parameters controlling both upper and lower dependence, more complex dependence structure can be modeled.

The two-parameter copula in Archimedean family can be defined through a generator function of the form

$$p_{k,\eta} = \left(\frac{1}{1+s^{\eta}}\right)^{k}, \qquad s \in [0,\infty), 0 < \eta \leq 1, k > 0. \tag{5.1}$$

In this family, one parameter k is used to capture lower tail dependence, defined as

$$\tau_L = \lim_{v \to 1^-} Pr(F_2(T_2) \geq v | F_1(T_1) \geq v),$$

and the other parameter η is used to characterize the upper tail dependence, defined as

$$\tau_U = \lim_{v \to 0^+} Pr(F_2(T_2) \leq v | F_1(T_1) \leq v).$$

Provided limits exist, $\tau_U \in [0, 1]$ and $\tau_L \in [0, 1]$, η and k have a one-to-one mapping with the lower and upper tail dependence that $\tau_L = 2^{-2k}$ and $\tau_U = 2 - 2^{2\eta}$.

Both Clayton and Gumbel copulas are two limiting scenarios of the two-parameter copula family. When $\eta = 1$, it degenerates to Clayton copula and when $k \to \infty$, it becomes a Gumbel copula.

In addition to extensions on the methodology part, a genome-wide analysis with the smoking variable (perhaps collapsed into two categories: ever smoke and never smoke) added (as an additional risk factor) to analyze the AREDS data would be worthwhile to perform.

5.2 ACKNOWLEDGMENT

This work was supported by the National Institutes of Health (EY024226 to principal investigator: Wei Chen). I would like to thank the AREDS investigators for collecting these valuable data and the International AMD Genomics Consortium for generating the genetic data and performing quality checks.

APPENDIX

EXACT ANALYTICAL DERIVATIVES FOR THE CLAYTON COPULA

Following, an exact analytical derivatives with respect to each margin for the Clayton Copula will be derived. The purpose of this part is to show that how analytically complex for just one type of copula function using exact form.

Clayton copula

Let $A(u, v, \eta) = u^{-\eta} + v^{-\eta} - 1$, then the copula density function w.r.t u, v is:

$$\begin{aligned} c(u, v; \eta) = \frac{C(u, v; \eta)}{\partial u \partial v} &= \frac{(u^{-\alpha} + v^{-\eta} - 1)^{-1/\eta}}{\partial u \partial v} \\ &= \frac{(1+\eta)(u \cdot v)^{-1-\eta}}{A(u, v, \eta)^{-1-\frac{1}{\eta}}} \\ &= (1+\eta)(u \cdot v)^{-1-\eta}(u^{-\eta} + v^{-\eta} - 1)^{-\frac{1}{\eta}-2}. \end{aligned}$$

Now let's look at the first derivatives of the function $c(u, v; \eta)$ w.r.t u and η.

First derivative of $c(u, v; \eta)$ w.r.t η

$$\begin{aligned} \frac{\partial c}{\partial \eta} &= (u, v)^{-\eta-1}(u^{-\eta} + v^{-\eta} - 1)^{-2-\frac{1}{\eta}} - (1+\eta)(uv)^{-\eta-1} ln(uv)(u^{-\eta} + u^{-\eta} - 1)^{-2-\frac{1}{\eta}} + \\ &\quad (1+\eta)(uv)^{-\eta-1}(u^{-\eta} + u^{-\eta} - 1)^{-2-\frac{1}{\eta}} \\ &\quad \left(\frac{ln(u^{-\eta} + u^{-\eta} - 1)}{\eta^2} + \frac{(-2-\frac{1}{\eta})(-u^{-\eta} ln(u) - v^{-\eta} ln(v))}{u^{-\eta} + u^{-\eta} - 1} \right). \end{aligned}$$

First derivative of $c(u, v; \eta)$ w.r.t u

$$\begin{aligned} \frac{\partial c}{\partial u} &= (1+\eta)(uv)^{-\eta-1}(-\eta-1)(u^{-\eta} + v^{-\eta} - 1)^{-2-\eta^{-1}} u^{-1} \\ &\quad - (1+\eta)\eta(-2-\frac{1}{\eta})(uv)^{-\eta-1} u^{-1-\eta}(u^{-\eta} + v^{-\eta} - 1)^{3-\eta^{-1}}. \end{aligned}$$

Using the fact:

$$\frac{\partial A}{\partial \eta} = -u^{-\eta} ln(u) - v^{-\eta} ln(v), \quad \frac{\partial^2 A}{\partial^2 \eta} = u^{-\eta} ln(u)^2 + v^{-\eta} ln(v)^2$$

and

$$\frac{\partial A}{\partial u} = -\eta u^{-\eta-1}, \quad \frac{\partial^2 A}{\partial^2 u} = \eta(\eta+1)u^{-\eta-2}.$$

We have following second derivatives:

$$\frac{\partial^2 c}{\partial^2 \eta} = \frac{\partial c}{\partial \eta} \cdot \left(-ln(v) + \frac{ln(A(u,v,\eta))}{\eta^2} + \frac{(-2-\frac{1}{\eta})\frac{\partial A}{\partial \eta}}{A(u,v,\eta)} \right) + c(u,v) \cdot \left(\frac{\frac{\frac{\partial A}{\partial \eta}}{A(u,v,\eta)}\eta^2 - 2ln(A(u,v,\eta))\eta}{\eta^4} + \frac{\left(\frac{1}{\eta^2}\frac{\partial A}{\partial \eta} + (-2-\frac{1}{\eta})\frac{\partial^2 A}{\partial^2 \eta}\right) \cdot A(u,v,\eta) - (-2-\frac{1}{\eta})\left(\frac{\partial A}{\partial \eta}\right)^2}{A(u,v,\eta)^2} \right)$$

$$\frac{\partial^2 c}{\partial^2 u} = \frac{\frac{\partial c}{\partial u}(\eta+1)u - (\eta+1)c(u,v)}{u^2} + \frac{(2+\frac{1}{\eta})\left(\frac{\partial c}{\partial u}\frac{\partial A}{\partial u} + c(u,v)\frac{\partial^2 A}{\partial^2 u}\right) - c(u,v)(2+\frac{1}{\eta})\left(\frac{\partial^2 A}{\partial^2 u}\right)^2}{A(u,v,\eta)^2}$$

$$\frac{\partial^2 c}{\partial u \eta} = \frac{\frac{\partial c}{\partial \eta}(\eta+1) + c(u,v)}{u} + \frac{(u^{\eta+1}A(u,v,\eta))\left[\frac{\partial c}{\partial \eta}(2\eta+1) + 2c(u,v)\right]}{u^{2\eta+2}A(u,v,\eta)^2} + \frac{c(u,v)(2\eta+1)\left[u^{\eta+1}ln(u)A(u,v,\eta) + u^{\eta+1}\frac{\partial A}{\partial \eta}\right]}{u^{2\eta+2}A(u,v,\eta)^2}$$

$$\frac{\partial^2 c}{\partial u v} = \frac{\frac{\partial c}{\partial v}(\eta+1)}{u} + \frac{\frac{\partial c}{\partial v}(2\eta+1)}{u^{2\eta+2}A(u,v,\eta)} - \frac{c(u,v)(2\eta+1)}{u^{\eta+2}A(u,v,\eta)^2} \times \frac{\partial A}{\partial v}.$$

With a specific marginal distribution function, we can derive complete analytical formulas for each Clayton copula model.

BIBLIOGRAPHY

Age-Related Eye Disease Study Research Group (1999) The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. *Control Clinical Trials*, **20**, 573–600.

Akaike, H. (1998) *Information Theory and an Extension of the Maximum Likelihood Principle*. Springer New York.

Azzato, E. M., Pharoah, P. D., Harrington, P., Easton, D. F., Greenberg, D., Caporaso, N. E., Chanock, S. J., Hoover, R. N., Thomas, G., Hunter, D. J. and Kraft, P. (2010) A genome-wide association study of prognosis in breast cancer. *Cancer Epidemiol Biomarkers Prev*, **19**, 1140–1143.

de Boor, C. (2011) *A Practical Guide to Splines (Applied Mathematical Sciences)*. Springer.

Breslow, N. (1972) Discussion of the paper by D. R. Cox. JR. *Journal of Royal Statististic Society. Series: B*, 216–217.

Cantor, R. M., Lange, K. and Sinsheimer, J. S. (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American Journal of Human Genetics*, **86**, 6–22.

Chen, H., Lumley, T., Brody, J., Heard-Costa, N. L., Fox, C. S., Cupples, L. A. and Dupuis, J. (2014) Sequence kernel association test for survival traits. *Genet Epidemiol*, **38**, 191–197.

Chen, W., Stambolian, D., Edwards, A. O., Branham, K. E., Othman, M. and al., J. J. (2010a) Genetic variants near TIMP3 and high-density lipoproteinassociated loci influence susceptibility to age-related macular degeneration. *PNAS*, **107**, 7401–7406.

Chen, X., Fan, Y., Pouzo, D. and Ying, Z. (2010b) Estimation and model selection of semiparametric multivariate survival functions under general censorship. *Journal of Econometrics*, **157(2)**, 129–142.

Chen, Z. (2012) *A Flexible Copula Model for Bivariate Survival Data*. Ph.D. thesis, University of Rochester.

Cho, Y. S., Go, M. J., Kim, Y. J. and Heo, J. Y. (2009) A large-scale genome-wide association study of asian populations uncovers genetic factors influencing eight quantitative traits. *Nature Genetics*, **41**, 527–534.

Clayton, D. G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141–151.

Cox, D. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, **34**, 187–220.

Cox, D. and Hinkley, D. (1979) *Theoretical Statistics*. Chapman and Hall/CRC.

Ding, Y., Liu, Y., Yan, Q., Fritsche, L. G., Cook, R. J., Clemons, T., Ratnapriya, R., Klein, M. L., Abecasis, G. R., Swaroop, A., Chew, E. Y., Weeks, D. E. and Chen, W. (2017) Bivariate analysis of age-related macular degeneration progression using genetic risk scores. *Genetics*, **206**, 119–133.

Ding, Y. and Nan, B. (2011) A sieve m-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Annals of Statistics*, **39**, 2795–3443.

Fan, R., Wang, Y., Mills, J. L., Carter, T. C., Lobach, I., Wilson, A. F., Bailey-Wilson, J. E., Weeks, D. E., and Xiong, M. (2014) Generalized functional linear models for case-control association studies. *Genet Epidemiol*, **38**, 622–637.

Fan, R., Wang, Y., Mills, J. L., Wilson, A. F., Bailey-Wilson, J. E., and Xiong, M. (2013) Functional linear models for association analysis of quantitative traits. *Genet Epidemiol*, **37**, 726 – 742.

Fan, R., Wang, Y., Yan, Q., Ding, Y., Weeks, D., Lu, Z., Ren, H., Cook, R. J., Xiong, M., Swaroop, A., Chew, E. Y., and Chen, W. (2016) Gene-based association analysis for censored traits via fixed effect functional regressions. *Genet Epidemiol*, **40**, 133–143.

Fritsche, L. G., Chen, W. and et al., M. S. (2013) Seven new loci associated with age-related macular degeneration. *Nature Genetics*, **45**, 433–439.

Fritsche, L. G., Igl, W., Baileyet, J. N. C. et al. (2016) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics*, **48**, 134–143.

Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. and Amos, C. I. (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet*, **82**, 100–112.

Gumbel, E. J. (1960) Bivariate exponential distributions. *Journal of the American Statistical Association*, **55**, 698–707.

Han, F. and Pan, W. (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, **70**, 537–545.

He, W. and Lawless, J. F. (2003) Flexible maximum likelihood methods for bivariate proportional hazards models. *Biometrics*, **59**, 837–848.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data*. Springer New York.

Ioannidis, J. P. A., Castaldi, P. and Evangelou, E. (2010) A compendium of genome-wide associations for cancer: critical synopsis and reappraisal. *Journal of the National Canser Institution*, **102**, 846–858.

Joe, H. (1997) *Multivariate models and dependence concepts*. Chapman & Hall.

Kim, G., Silvapulle, M. J. and Silvapulle, P. (2007) Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis.*, **51**, 2836–2850.

Klein, R., Chou, C.-F., Klein, B. E. K. et al. (2011) Prevalence of age-related macular degeneration in the us population. *Archives of Ophthalmology*, **129**, 75–80.

Lawless, J. F. and Yilmaz, Y. E. (2011) Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*, **53**, 779–796.

Lee, E. W., Wei, L. J. and Amato, D. A. (1992) Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival analysis: State of the art*, **211**, 237–247.

Li, B. and Leal, S. M. (2008) Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*, **83**, 311–321.

Liang, K. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lin, D.-Y. and Tang, Z.-Z. (2011) A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet*, **89**, 354–367.

Lin, X., Cai, T., Wu, M., Zhou, Q., Liu, G., Christiani, D. C. and Lin, X. (2011) Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology*, **35**, 620–631.

Oakes, D. (1982) A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B*, **44**, 414–422.

Pillas, D., Hoggart, C. J. and Evans, D. M. (2010) Genome-wide association study reveals multiple loci associated with primary tooth development during infancy. *PLOS Genetics*, **6**, e1000856.

Ramsay, J. O., Hooker, G. and Graves, S. (2009) *Functional Data Analysis With R and Matlab*. Springer New York.

Richardson, L. F. (1911) The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *The Royal Society A*, **210**, 307–357.

Sardell, R. J., Persad, P. J., Pan, S. S. and et al., P. W. (2016) Progression rate from intermediate to advanced Age-Related Macular Degeneration is correlated with the wumber of risk alleles at the CFH locus. *Invest Ophthalmol Vis Sci*, **57**, 6107–6115.

Schaffner, S. F., Foo, C., Gabriel1, S., Reich, D., Daly1, M. J. and Altshuler, D. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*, **15**, 1576–1583.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Seddon, J. M., Reynolds, R., Maller, J., Fagerness, J. A., Daly, M. J. and Rosner, B. (2009) Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Investigative Ophthalmology & Visual Science*, **50**, 2044–53.

Seddon, J. M., Reynolds, R., Yu, Y. and Rosner, B. (2014) Three new genetic loci (R1210C in CFH variants in COL8A1 and RAD51B) are independently related to progression to advanced macular degeneration. *PLoS ONE*, **9(1)**, 1–11.

Sha, Q., Zhang, Z. and Zhang, S. (2011) An improved score test for genetic association studies. *Genetic Epidemiology*, **35**, 350–9.

Shih, J. H. and Louis, T. A. (1995) Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, **51**, 1384–1399.

Sklar, A. (1959) Fonctions de repartition a n dimensions et leurs marges. *Publications de L Institut de Statistique de L Universite de Paris*, **8**, 229–231.

Swaroop, A., Chew, E. Y., Abecasis, G. R. et al. (2009) Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration. *Annual review of genomics and human genetics*, **10**, 19–43.

The Eye Diseases Prevalence Research Group (2004) Causes and Prevalence of Visual Impairment Among Adults in the United States. *Archives of Ophthalmology*, **122**, 477–485.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Vaupel, J., Manton, K. and Stallard, E. (1979) The impact of heterogeneity in individual frailty and the dynamics of mortality. *Demography*, **16**, 439–454.

Voorman, A., Brody, J., Chen, H., Lumley, T. and Davis, B. (2016) *cluster: Meta-Analysis of Region-Based Tests of Rare DNA Variants*.

Wang, W. and Wells, M. T. (2000) Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association*, **95**, 62–72.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G. and Todd, J. A. (2005) Genome-wide association studies: theoretical and practical concernss. *Nature Reviews Genetics*, **6**, 109–118.

Wei, L. (1992) The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Journal of the Royal Statistical Society: Series B*, **11**, 1871–1879.

Wei, L. J., Lin, D. and Weissfeld, L. (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, **84**, 1065–1073.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010) Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *American Journal of Human Genetics*, **11**, 929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, **89**, 82–93.

Zhou, Q., Hu, T. and Sun, J. (2017) A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association*, **112**, 664–672.